



Agilent GeneSpring GX

User Manual



Agilent Technologies

Notices

© Agilent Technologies, Inc. 2009

No part of this manual may be reproduced in any form or by any means (including electronic storage and retrieval or translation into a foreign language) without prior agreement and written consent from Agilent Technologies, Inc. as governed by United States and international copyright laws.

Edition

Revised November 2009

Printed in USA

Agilent Technologies, Inc.
3501 Stevens Creek Blvd.
Santa Clara, CA 95052 USA

Java™ is a U.S. trademark of Sun Microsystems, Inc.

Windows ® is a U.S. registered trademark of Microsoft Corporation.

Software Revision

This guide is valid for 11.0 revisions of the Agilent GeneSpring GX software.

Software Revision

The material contained in this document is provided “as is,” and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Technology Licenses

The hardware and/or software described in this document are furnished under a license and may be used or copied only in accordance with the terms of such license.

Restricted Rights Legend

U.S. Government Restricted Rights. Software and technical data rights granted to the federal government include only those rights customarily provided to end user customers. Agilent provides this customary commercial license in Software and technical data pursuant to FAR 12.211 (Technical Data) and 12.212 (Computer Software) and, for the Department of Defense, DFARS 252.227-7015 (Technical Data - Commercial Items) and DFARS 227.7202-3 (Rights in Commercial Computer Software or Computer Software Documentation).

Contents

Contents	i
List of Figures	iii
List of Tables	v
1 GeneSpring GX Installation	1
1.1 Supported and Tested Platforms	1
1.1.1 System Requirements for Copy Number and Association Experiments	2
1.1.2 Installation and Usage Requirements	2
1.1.3 GeneSpring GX Installation Procedure for Microsoft Windows	3
1.1.4 Activating your GeneSpring GX	4
1.1.5 Uninstalling GeneSpring GX from Windows	5
1.2 Installation on Linux	5
1.2.1 Installation and Usage Requirements	6
1.2.2 GeneSpring GX Installation Procedure for Linux	6
1.2.3 Activating your GeneSpring GX	7

1.2.4	Uninstalling GeneSpring GX from Linux	8
1.3	Installation on Apple Macintosh	8
1.3.1	Installation and Usage Requirements	9
1.3.2	GeneSpring GX Installation Procedure for Macintosh	9
1.3.3	Activating your GeneSpring GX	10
1.3.4	Uninstalling GeneSpring GX from Mac	12
1.4	License Manager	12
1.4.1	Utilities of the License Manager	13
1.5	Upgrade	15
2	GeneSpring GX Quick Tour	17
2.1	Introduction	17
2.2	Launching GeneSpring GX	17
2.3	GeneSpring GX User Interface	17
2.3.1	GeneSpring GX Desktop	18
2.3.2	Project Navigator	19
2.3.3	The Workflow Browser	19
2.3.4	Global lists	20
2.3.5	The Legend Window	20
2.3.6	Status Line	20
2.4	Organizational Elements and Terminology in GeneSpring GX	21

2.4.1	Project	21
2.4.2	Experiment	22
2.4.3	Sample	23
2.4.4	Experiment Grouping, Parameters and Parameter Values	23
2.4.5	Conditions and Interpretations	23
2.4.6	Entity List	27
2.4.7	Entity Tree, Condition Tree, Combined Tree and Classification	28
2.4.8	Class Prediction Model	29
2.4.9	Script	30
2.4.10	Pathway	30
2.4.11	Inspectors	30
2.4.12	Hierarchy of objects	32
2.4.13	Right-click operations	32
2.4.14	Search	36
2.4.15	Saving and Sharing Projects	39
2.4.16	Software Organization	39
2.5	Exporting and Printing Images and Reports	40
2.6	Scripting	40
2.7	Options	40
2.8	Update Utility	41
2.8.1	Product Updates	41

2.9	Getting Help	41
3	Technology and Biological Genome	43
3.1	Technology	43
3.1.1	Standard Technology Creation	44
3.1.2	Agilent eArray Technology Creation	45
3.1.3	Custom Technology Creation	47
3.1.4	Technology creation on the fly	47
3.1.5	Inspection of Technology	47
3.1.6	Technology Deletion	48
3.2	Update Technology	48
3.2.1	Automatic Query of Update Server	48
3.2.2	Update Technology Annotations	48
3.3	Translation	51
3.3.1	Implementation	55
3.3.2	Explicit Translation mapping	56
3.3.3	Translation in Copy number and Association experiments	59
3.4	Biological Genome	59
4	Data Migration	61
4.1	GeneSpring GX Data Migration from GeneSpring GX 7	61
4.1.1	Migrations Steps	61

4.1.2	Migrated Objects	65
4.2	Data Migration from WG5.2 to WG11	66
4.2.1	Users and Groups	66
4.2.2	Samples	67
4.2.3	Genomes, Projects, Experiments	68
4.2.4	Entity Lists, Gene Trees, Condition Trees and Classifications	69
4.2.5	Ownership and Permissions	69
4.2.6	Potential causes of Migration failure and Known Issues	70
4.3	Migration of GX11 Desktop Data to GX11 Workgroup	70
4.4	Migration of GeneSpring GX 10.0 to GeneSpring GX 11.0	71
5	Data Visualization	73
5.1	View	73
5.1.1	The View Framework in GeneSpring GX	73
5.1.2	View Operations	74
5.2	The Spreadsheet View	80
5.2.1	Spreadsheet Operations	81
5.2.2	Spreadsheet Properties	83
5.3	MvA plot	85
5.4	The Scatter Plot	85
5.4.1	Scatter Plot Operations	86

5.4.2	Scatter Plot Properties	88
5.5	The Profile Plot View	94
5.5.1	Profile Plot Operations	95
5.5.2	Profile Plot Properties	95
5.6	The Heatmap View	98
5.6.1	Heatmap Operations	98
5.6.2	Heatmap Toolbar	102
5.6.3	heatmap Properties	103
5.6.4	Heatmap for viewing Copy Number Analysis Results	104
5.7	The Histogram View	104
5.7.1	Histogram Operations	107
5.7.2	Histogram Properties	107
5.8	The Bar Chart	110
5.8.1	Bar Chart Operations	110
5.8.2	Bar Chart Properties	111
5.9	The Matrix Plot View	113
5.9.1	Matrix Plot Operations	114
5.9.2	Matrix Plot Properties	114
5.10	Summary Statistics View	117
5.10.1	Summary Statistics Operations	118
5.10.2	Summary Statistics Properties	118

5.11	The Box Whisker Plot	121
5.11.1	Box Whisker Operations	122
5.11.2	Box Whisker Properties	122
5.12	The Venn Diagram	126
5.12.1	Venn Diagram Operations	126
5.12.2	Venn Diagram Properties	129
5.13	LD Plot	130
5.13.1	LD Plot Toolbar	130
5.13.2	LD Measure Options	131
5.13.3	LD Plot Properties	132
5.14	Haplotypes view	132
5.14.1	Haplotypes Context Menu	133
5.15	Genome Browser	133
5.16	Plot Options	134
5.16.1	Plot Log10/Linear Values	134
5.16.2	Plot List Associated Values	134
5.17	Miscellaneous operations	135
5.17.1	Save Current view	135
5.17.2	Find Entity	136
5.17.3	Inspect Entities	136
5.17.4	Properties	137

6	Analyzing Affymetrix Expression Data	139
6.1	Running the Affymetrix Workflow	139
6.2	Data Processing	142
6.3	Guided Workflow steps	145
6.4	Advanced Workflow	160
6.4.1	Creating an Affymetrix Expression Experiment	160
6.4.2	Experiment Setup	169
6.4.3	Quality Control	173
6.4.4	Analysis	177
6.4.5	Class Prediction	180
6.4.6	Results	180
6.4.7	Utilities	181
6.4.8	Affymetrix Technology creation using Custom CDF	181
7	Affymetrix Summarization Algorithms	185
7.0.1	Probe Summarization Algorithms	185
7.0.2	Computing Absolute Calls	189
8	Analyzing Affymetrix Exon Expression Data	191
8.1	Running the Affymetrix Exon Workflow	191
8.2	Data Processing	194
8.3	Guided Workflow steps	197

8.4	Advanced Workflow	210
8.4.1	Creating an Affymetrix ExonExpression Experiment	211
8.4.2	Experiment setup	218
8.4.3	Quality Control	221
8.4.4	Analysis	224
8.4.5	Class Prediction	224
8.4.6	Results	224
8.4.7	Utilities	225
8.4.8	Algorithm Technical Details	225
9	Analyzing Affymetrix Exon Splicing Data	227
9.1	Running the Affymetrix Exon Splicing Workflow	227
9.1.1	Creating an Affymetrix Exon Splicing Experiment	228
9.1.2	Data Processing for Exon arrays	235
9.1.3	Experiment setup	236
9.1.4	Quality Control	239
9.1.5	Analysis	242
9.1.6	Exon Splicing Analysis	243
9.1.7	Class Prediction	255
9.1.8	Results	257
9.1.9	Utilities	257

9.1.10	Algorithm Technical Details	258
9.2	Tutorial for Exon Splicing Analysis	258
10	Analyzing Illumina Data	263
10.1	Running the Illumina Workflow:	263
10.2	Data Processing for Illumina arrays	266
10.3	Guided Workflow steps	268
10.4	Advanced Workflow:	282
10.4.1	Experiment Setup	287
10.4.2	Quality control	291
10.4.3	Analysis	295
10.4.4	Class Prediction	296
10.4.5	Results	298
10.4.6	Utilities	298
10.4.7	Illumina Custom Technology creation	299
11	Analyzing Agilent Single Color Expression Data	301
11.1	Running the Agilent Single Color Workflow	301
11.1.1	Analyzing Agilent Two Color data in Agilent Single Color Experiment Type	304
11.2	Data Processing for Agilent Single Color arrays	307
11.3	Guided Workflow steps	309
11.4	Advanced Workflow	323

11.4.1	Experiment Setup	328
11.4.2	Quality Control	331
11.4.3	Analysis	336
11.4.4	Class Prediction	337
11.4.5	Results	338
11.4.6	Utilities	339
12	Analyzing Agilent Two Color Expression Data	341
12.1	Running the Agilent Two Color Workflow	341
12.2	Data Processing for Agilent Two Color arrays	347
12.3	Guided Workflow steps	348
12.4	Advanced Workflow	361
12.4.1	Experiment Setup	364
12.4.2	Quality Control	369
12.4.3	Analysis	373
12.4.4	Class Prediction	374
12.4.5	Results	374
12.4.6	Utilities	377
12.5	Custom Agilent Arrays	378
13	Analyzing Agilent miRNA Data	381
13.1	Running the Agilent miRNA Workflow	381

13.1.1	Sample validation in GeneSpring GX 11.0	387
13.2	Data Processing	387
13.3	Guided Workflow steps	388
13.3.1	Summary Report (Step 1 of 8)	388
13.3.2	Experiment Grouping (Step 2 of 8)	388
13.3.3	Quality Control (QC) (Step 3 of 8)	390
13.3.4	Filter probesets (Step 4 of 8)	394
13.3.5	Significance Analysis (Step 5 of 8)	394
13.3.6	Fold-change (Step 6 of 8)	398
13.3.7	Gene Ontology Analysis (Step 7 of 8)	400
13.3.8	Find Significant Pathways (Step 8 of 8)	403
13.4	Advanced Workflow	404
13.4.1	Experiment Setup	408
13.4.2	Quality Control	409
13.4.3	Analysis	412
13.4.4	Class Prediction	416
13.4.5	Results	416
13.4.6	TargetScan	416
13.4.7	Utilities	419
14	Analyzing Real Time PCR Data	421

14.1	Running the Real Time PCR Workflow	421
14.1.1	Technology Creation in RT-PCR experiments	425
14.1.2	Data Processing	425
14.1.3	Experiment Setup	426
14.1.4	Quality Control	426
14.1.5	Analysis	428
14.1.6	Class Prediction	428
14.1.7	Results	429
14.1.8	Utilities	429
15	Analyzing Generic Single Color Expression Data	433
15.1	Creating Technology	433
15.1.1	Project and Experiment Creation	443
15.2	Data Processing for Generic Single Color Experiment	444
15.3	Advanced Analysis	446
15.3.1	Experiment Setup	450
15.3.2	Quality Control	453
15.3.3	Analysis	457
15.3.4	Class Prediction	460
15.3.5	Results	460
15.3.6	Utilities	460

16 Analyzing Generic Two Color Expression Data	461
16.1 Creating Technology	461
16.1.1 Creation of Custom Technology-Non gpr files	461
16.1.2 GenePix Result Technology creation	467
16.1.3 Project and Experiment Creation	468
16.2 Advanced Analysis	471
16.2.1 Data Processing for Generic Two Color Data	476
16.2.2 Experiment Setup	477
16.2.3 Quality Control	480
16.2.4 Analysis	483
16.2.5 Class Prediction	486
16.2.6 Results	486
16.2.7 Utilities	487
17 Loading Experiment from NCBI GEO	489
17.1 Introduction	489
17.1.1 Load a GSE dataset	489
17.1.2 Experiment Parameters	491
17.2 Possible Error Messages	492
17.3 Experiment Parameters and Sample Attributes	495
17.3.1 Create Experiment Parameters from Sample Attributes	495

18 Advanced Workflow	497
18.1 Experiment Setup	497
18.1.1 Quick Start Guide	497
18.1.2 Experiment Grouping	498
18.1.3 Create Interpretation	499
18.1.4 Create new Gene Level Experiment	502
18.2 Quality Control	506
18.2.1 Quality Control on Samples	506
18.2.2 Filter Probesets by Expression	507
18.2.3 Filter probesets by Flags	509
18.2.4 Filter Probesets on Data Files	511
18.2.5 Filter Probesets by Error	511
18.3 Analysis	512
18.3.1 Statistical Analysis	512
18.3.2 Filter on Volcano Plot	522
18.3.3 Fold change	524
18.3.4 Clustering	528
18.3.5 Find similar entities	528
18.3.6 Filter on Parameters	529
18.3.7 Principal Component Analysis	531
18.4 Class Prediction	535

18.4.1	Build Prediction model	535
18.4.2	Run prediction	535
18.5	Results Interpretation	536
18.5.1	GO Analysis	538
18.5.2	GSEA	538
18.6	Find Similar Objects	538
18.6.1	Find Similar Entity lists	538
18.6.2	Find Similar Pathways	539
18.7	Utilities	540
18.7.1	Save Current view	540
18.7.2	Genome Browser	540
18.7.3	Import Entity List from file	540
18.7.4	Import BROAD GSEA Genesets	541
18.7.5	Import BIOPAX pathways	541
18.7.6	Differential Expression Guided Workflow	541
18.7.7	Filter on Entity List	541
19	Normalization, Statistical Hypothesis Testing, and Differential Expression Analysis	545
19.1	Threshold	545
19.2	Normalization Algorithms	545
19.2.1	Percentile Shift Normalization	546

19.2.2	Scale	546
19.2.3	Quantile Normalization	546
19.2.4	Normalize to control genes	547
19.2.5	Normalize to External Value	547
19.2.6	Lowess Normalization	548
19.3	Details of Statistical Tests in GeneSpring GX	549
19.3.1	The Unpaired <i>t</i> -Test for Two Groups	549
19.3.2	The <i>t</i> -Test against 0 for a Single Group	549
19.3.3	The Paired <i>t</i> -Test for Two Groups	549
19.3.4	The Unpaired Unequal Variance <i>t</i> -Test (Welch <i>t</i> -test) for Two Groups	550
19.3.5	The Unpaired Mann-Whitney Test	550
19.3.6	The Paired Mann-Whitney Test	550
19.3.7	One-Way ANOVA	551
19.3.8	Post hoc testing of ANOVA results	552
19.3.9	Unequal variance (Welch) ANOVA	553
19.3.10	The Kruskal-Wallis Test	553
19.3.11	The Repeated Measures ANOVA	554
19.3.12	The Repeated Measures Friedman Test	554
19.3.13	The N-way ANOVA	555
19.4	Obtaining <i>p</i> -Values	556
19.4.1	<i>p</i> -values via Permutation Tests	557

19.5	Adjusting for Multiple Comparisons	557
19.5.1	Bonferroni	558
19.5.2	Bonferroni Step-down (Holm method)	558
19.5.3	The Westfall-Young method	559
19.5.4	The Benjamini-Hochberg method	559
19.5.5	The Benjamini-Yekutieli method	560
19.5.6	Recommendations	560
19.5.7	FAQ	560
20	Clustering: Identifying Genes and Conditions with Similar Expression Profiles with Similar Behavior	563
20.1	What is Clustering	563
20.2	Clustering Wizard	564
20.3	Graphical Views of Clustering Analysis Output	566
20.3.1	Cluster Set or Classification	567
20.3.2	Dendrogram	570
20.3.3	U Matrix	577
20.4	Distance Measures	578
20.5	K-Means	580
20.6	Hierarchical	580
20.7	Self Organizing Maps (SOM)	582
20.8	Missing Value Handling	583

21 Class Prediction: Learning and Predicting Outcomes	585
21.1 General Principles of Building a Prediction Model	585
21.2 Prediction Pipeline	586
21.2.1 Validate	586
21.2.2 Prediction Model	588
21.3 Running Class Prediction in GeneSpring GX	588
21.3.1 Build Prediction Model	588
21.3.2 Run Prediction	590
21.4 Decision Trees	591
21.4.1 Decision Tree Model Parameters	594
21.4.2 Decision Tree Model	595
21.5 Neural Network	596
21.5.1 Neural Network Model Parameters	597
21.5.2 Neural Network Model	597
21.6 Support Vector Machines	599
21.6.1 SVM ModelParameters	600
21.7 Naive Bayesian	602
21.7.1 Naive Bayesian Model Parameters	602
21.7.2 Naive Bayesian Model View	603
21.8 Partial Least Square Discrimination	603
21.8.1 PLSD Model and Parameters	604

21.9	Viewing Classification Results	605
21.9.1	Confusion Matrix	605
21.9.2	Classification Report	606
21.9.3	Lorenz Curve	606
22	Gene Ontology Analysis	609
22.1	Working with Gene Ontology Terms	609
22.2	Introduction to GO Analysis in GeneSpring GX	611
22.3	GO Analysis	611
22.4	GO Analysis Views	614
22.4.1	GO Spreadsheet	614
22.4.2	The GO Tree View	614
22.4.3	The Pie Chart	616
22.5	GO Enrichment Score Computation	620
23	Gene Set Enrichment Analysis	623
23.1	Introduction to GSEA	623
23.2	Gene sets	623
23.3	Performing GSEA in GeneSpring GX	624
23.4	GSEA Computation	628
23.5	Import BROAD GSEA Genesets	630
24	Gene Set Analysis	631

24.1 Introduction to GSA	631
24.2 Gene sets	631
24.3 Performing GSA in GeneSpring GX	632
24.4 GSA Computation	636
25 Pathway Analysis	639
25.1 Introduction to Pathway Analysis	639
25.2 Licensing	640
25.3 Getting Started	640
25.4 Working with Other Organisms	641
25.5 Pathway Analysis in Microarray Experiment	641
25.5.1 Pathways, Entities and Relationships	642
25.5.2 Analysis	642
25.5.3 Pathway View	652
25.5.4 Layouts	666
25.5.5 Themes	668
25.6 Extract Relations via NLP	668
25.6.1 NLP Settings	672
25.7 MeSH Pathway Builder	673
25.7.1 Launching MeSH Pathway Builder	674
25.8 Find Significant Pathways	676

25.8.1	The BioPAX format	676
25.8.2	Prepackaged Pathways and Migrating Older Pathways	679
25.8.3	Import from PathwayArchitect	680
25.8.4	Find Significant Pathways	681
25.9	Pathway Experiment	683
25.9.1	Launching a Pathway Experiment	683
25.9.2	Lassoing	687
25.9.3	Simple Analysis	689
25.9.4	Advanced Analysis	690
25.9.5	Exporting Pathways	690
25.10	Pathway Database	691
25.10.1	Pathway Database Organization Overview	692
25.10.2	Database Entities	693
25.10.3	Relations	696
25.10.4	Database statistics	700
25.10.5	Overview of Natural Language Processing (NLP)	702
25.11	Update Pathway Interactions	703
25.12	Working with the Pathway Interactions Server	703
25.13	Troubleshooting	704

26 Copy Number Analysis

707

26.1	Introduction	707
26.1.1	Terminology in Copy Number analysis	708
26.2	Technologies supported by GeneSpring GX 11.0	708
26.2.1	Experiment Creation	708
26.2.2	Reference	710
26.2.3	Copy Number Analysis	712
26.2.4	Special mention for Affymetrix Mapping 100k Array	716
26.3	Workflow description for Affymetrix files	717
26.3.1	Create Technology	717
26.3.2	Creating a Copy Number experiment	718
26.3.3	Experiment Setup	721
26.3.4	Quality Control	721
26.3.5	Copy Number Analysis	727
26.3.6	Common Genomic Variant Regions	729
26.3.7	Filters	733
26.3.8	Views	739
26.3.9	Results Analysis	741
26.3.10	Utilities	741
26.4	Copy Number analysis of Illumina	743
26.4.1	Obtaining Data from Illumina	743
26.4.2	Handling Missing Values	745

26.4.3	Workflow description for Illumina Outputs	745
26.5	Create Custom Reference	745
26.6	Useful information	747
26.6.1	Using disc cache	747
26.6.2	Entity Lists and Translation rules in Copy Number	747
26.6.3	Configuration options for Copy Number analysis	747
26.6.4	Performance Statistics for Copy Number Analysis	749
26.7	Copy Number Algorithms	749
26.7.1	BRLMM	749
26.7.2	Hidden Markov Model (HMM)	750
26.7.3	Canary algorithm	754
26.7.4	Birdseed algorithm	754
26.7.5	CBS for segmenting genome with respect to Copy Number	757
26.7.6	Post Processing to assign Copy Numbers to segments created by CBS	757
26.7.7	Fawkes algorithm	758
26.8	Tutorials for Copy Number Analysis	760
27	Association Analysis	761
27.1	Introduction	761
27.2	Technology	762
27.3	Experiment Creation	762

27.3.1	Illumina Association Analysis Experiment	763
27.3.2	Affymetrix Association Analysis Experiment	765
27.4	Quality Control	765
27.4.1	Filter Samples by Missing Values	766
27.4.2	Birdseed Report	767
27.4.3	EIGENSTRAT Filter on Samples	768
27.5	Filters	772
27.5.1	Filter SNPs by Missing Value	772
27.5.2	Identify SNPs with Differential Missingness	773
27.5.3	Filter SNPs by HWE p-value	775
27.5.4	Filter SNPs by MAF	776
27.6	Analysis	777
27.6.1	EIGENSTRAT Correction on Samples	777
27.6.2	Statistical Analysis	782
27.6.3	SNP Tagging	790
27.6.4	SNP Regression	792
27.6.5	Haplotype Trend Regression	793
27.6.6	LD Analysis	796
27.7	Views	798
27.7.1	Genome Browser	798
27.8	Results Interpretations	800

27.8.1	Identify Overlapping Genes	800
27.9	Utilities	800
27.9.1	Using disc cache	801
28	The Genome Browser	803
28.1	Introduction	803
28.2	Tracks in Genome Browser	804
28.2.1	Track functionalities	804
28.3	Visualization in Genome Browser	805
28.4	Working with Genome Browser	807
28.4.1	Manage Genome Browser Data	809
28.4.2	Drag and Drop Experiments	813
28.4.3	Drag and Drop Entity Lists	815
28.4.4	Drag and Drop Files from anywhere	815
28.4.5	Track Operations	817
28.4.6	Track properties	817
28.5	Viewing Copy Number Experiments in Genome Browser	819
28.5.1	Data columns	819
28.5.2	Utilities for Copy Number Experiments	820
28.6	Useful details to know	820
28.7	FAQ	821

29 Ingenuity Pathways Analysis (IPA) Connector	823
29.1 Using the GeneSpring GX -IPA Connector	823
29.1.1 Create Pathway in IPA	823
29.1.2 Import List from IPA	826
29.1.3 Perform Data Analysis on Experiment	832
29.1.4 Perform Data Analysis on Entity List	839
30 GeneSpring GX Workgroup Client	849
30.1 Users and Groups	849
30.1.1 Login	850
30.2 Operations on GeneSpring Objects	850
30.2.1 Object ownership	851
30.2.2 Object permissions	851
30.2.3 Conflicts with permissions	852
30.2.4 Propagating permissions	852
30.2.5 Inheriting Permissions	854
30.3 Remote Execution	854
30.3.1 Task Manager	854
30.3.2 Remotely Executable Operations	855
30.3.3 Interpreting Task Logs	857
31 Writing Scripts in GeneSpring GX	859

31.1	The Script Editor	859
31.2	Hierarchy of data organization in GeneSpring GX	860
31.2.1	Accessing Projects, Experiments and their Constituent Elements	860
31.2.2	Accessing the Experiment Dataset	862
31.2.3	Some More Useful Functions	864
31.2.4	Some Common Marks	867
31.2.5	Creating UI Components	868
31.2.6	Example Scripts	870
31.3	The R Editor	877
31.3.1	Commands related to R input from GeneSpring GX	877
31.3.2	Commands related to R output to GeneSpring GX	878
31.3.3	Debugging a Script	879
31.3.4	Example R scripts	880
32	Table of Key Bindings and Mouse Clicks	885
32.1	Mouse Clicks and their actions	885
32.1.1	Global Mouse Clicks and their actions	885
32.1.2	Some View Specific Mouse Clicks and their Actions	886
32.1.3	Mouse Click Mappings for Mac	886
32.2	Key Bindings	886
32.2.1	Global Key Bindings	886

List of Figures

1.1	Activation Failure	5
1.2	Activation Failure	8
1.3	Activation Failure	11
1.4	The License Description Dialog	12
1.5	Confirm Surrender Dialog	14
1.6	Manual Surrender Dialog	14
1.7	Change License Dialog	15
1.8	License Re-activation Dialog	15
2.1	GeneSpring GX Layout	18
2.2	The Workflow Window	19
2.3	The Legend Window	20
2.4	Status Line	20
2.5	Confirmation Dialog	41
2.6	Product Update Dialog	42

3.1	Create Technology	43
3.2	Technology Creation	46
3.3	Technology Update	46
3.4	Data Library Updates Dialog	49
3.5	Automatic Download Confirmation Dialog	49
3.6	Update Technology Annotations	50
3.7	Input Parameters	51
3.8	Format data file	52
3.9	Choose Annotation Columns	53
3.10	Input Parameters	56
3.11	Translation Table	57
3.12	Save Entity List	58
3.13	Create Biological Genome	60
4.1	Experiment Exporter	62
4.2	Confirmation Window	62
4.3	Migrate GS7 Data	63
4.4	Partially Migrated Genomes	64
5.1	Export submenus	76
5.2	Export Image Dialog	76
5.3	Tools →Options Dialog for Export as Image	77

5.4	Error Dialog on Image Export	78
5.5	Menu accessible by Right-Click on the plot views	78
5.6	Menu accessible by Right-Click on the table views	80
5.7	Spreadsheet	81
5.8	Spreadsheet Properties Dialog	82
5.9	MvA plot	86
5.10	Scatter Plot	87
5.11	Scatter Plot Properties	88
5.12	Viewing Profiles and Error Bars using Scatter Plot	91
5.13	Scatter plot with Fold Change lines	92
5.14	Profile Plot	93
5.15	Profile Plot Properties	96
5.16	Heat Map	99
5.17	Export submenus	100
5.18	Export Image Dialog	100
5.19	Error Dialog on Image Export	101
5.20	heatmap Toolbar	102
5.21	heatmap Properties	103
5.22	Histogram	105
5.23	Histogram Viewing Options	106
5.24	Histogram Properties	108

5.25	Bar Chart	110
5.26	Matrix Plot	113
5.27	Matrix Plot Properties	115
5.28	Summary Statistics View	118
5.29	Summary Statistics Properties	119
5.30	Box Whisker Plot	121
5.31	Box Whisker Properties	123
5.32	The Venn Diagram	127
5.33	Create New Entity List from Venn Diagram	128
5.34	The Venn Diagram Properties	129
5.35	LD Plot Toolbar	131
5.36	Plot List Associated Values	135
5.37	Plot List Associated Values-Scatter plot	136
5.38	Plot List Associated Values-Profile plot	137
5.39	Plot List Associated Values-Histogram	138
6.1	Welcome Screen	140
6.2	Create New project	140
6.3	Experiment Selection	141
6.4	Experiment Description	143
6.5	Load Data	143

6.6	Choose Samples	144
6.7	Reordering Samples	144
6.8	Summary Report	146
6.9	Experiment Grouping	148
6.10	Edit or Delete of Parameters	149
6.11	Quality Control on Samples	150
6.12	Filter Probesets-Single Parameter	151
6.13	Filter Probesets-Two Parameters	152
6.14	Rerun Filter	152
6.15	Significance Analysis-T Test	155
6.16	Significance Analysis-Anova	156
6.17	Fold Change	157
6.18	GO Analysis	158
6.19	Find Significant Pathways	159
6.20	Load Data	162
6.21	Choose Technology and Template	162
6.22	Select Row Scope for Import	163
6.23	Choose Identifier and Signal Column	164
6.24	Single Colour Many Samples in one File Selection	165
6.25	Select ARR files	166
6.26	Summarization Algorithm	167

6.27	Normalization and Baseline Transformation	168
6.28	Normalize to control genes	169
6.29	Baseline Transformation	170
6.30	Gene Level Experiment Creation	171
6.31	Gene Level Experiment Creation - Normalization Options	172
6.32	Gene Level Experiment Creation - Choose Entities	173
6.33	Gene Level Experiment Creation - Preprocess Baseline Options	174
6.34	Quality Control	175
6.35	Entity list and Interpretation	176
6.36	Input Parameters	177
6.37	Output Views of Filter by Flags	178
6.38	Save Entity List	179
6.39	Confirmation Dialog Box	182
6.40	Choose Input Files	183
8.1	Welcome Screen	192
8.2	Create New project	192
8.3	Experiment Selection	193
8.4	Experiment Description	195
8.5	Load Data	195
8.6	Choose Samples	196

8.7	Reordering Samples	196
8.8	Summary Report	198
8.9	Experiment Grouping	200
8.10	Edit or Delete of Parameters	201
8.11	Quality Control on Samples	202
8.12	Filter Probesets-Single Parameter	203
8.13	Filter Probesets-Two Parameters	203
8.14	Rerun Filter	204
8.15	Significance Analysis-T Test	207
8.16	Significance Analysis-Anova	208
8.17	Fold Change	208
8.18	GO Analysis	209
8.19	Find Significant Pathways	210
8.20	Load Data	212
8.21	Select ARR files	212
8.22	Summarization Algorithm	214
8.23	Normalization	215
8.24	Search entities	216
8.25	Output Views	216
8.26	Choose Entities	217
8.27	Normalization and Baseline Transformation	217

8.28	Gene Level Experiment Creation	219
8.29	Gene Level Experiment Creation - Normalization Options	220
8.30	Gene Level Experiment Creation - Choose Entities	221
8.31	Gene Level Experiment Creation - Preprocess Baseline Options	222
8.32	Quality Control	223
9.1	Load Data	229
9.2	Error Message	229
9.3	Select ARR files	230
9.4	Pairing of CHP files	231
9.5	Summarization Algorithm	233
9.6	Normalization	234
9.7	Normalize to control genes	234
9.8	Normalization and Baseline Transformation	235
9.9	Gene Level Experiment Creation	238
9.10	Gene Level Experiment Creation - Normalization Options	239
9.11	Gene Level Experiment Creation - Choose Entities	240
9.12	Gene Level Experiment Creation - Preprocess Baseline Options	241
9.13	Quality Control	242
9.14	Input Data	243
9.15	Filtering Options	244

9.16	Output Views	245
9.17	Save Entity List	246
9.18	Input Data	246
9.19	Filtering of Probesets	247
9.20	Multiple Testing Correction	248
9.21	Results	249
9.22	Save Entity List	250
9.23	Input Data	251
9.24	Pairing Options	251
9.25	Results	252
9.26	Save Entity List	253
9.27	Input Data	254
9.28	Visualization	254
9.29	Visualization	255
9.30	Save Entity List	256
9.31	Gene Normalized Variance Plot	261
9.32	Gene Normalized Profile Plot	262
10.1	Welcome Screen	264
10.2	Create New project	264
10.3	Experiment Selection	265

10.4 Experiment Description	267
10.5 Load Data	267
10.6 Choose Samples	268
10.7 Summary Report	269
10.8 Experiment Grouping	271
10.9 Edit or Delete of Parameters	272
10.10Quality Control on Samples	273
10.11Filter Probesets-Single Parameter	274
10.12Filter Probesets-Two Parameters	275
10.13Rerun Filter	275
10.14Significance Analysis-T Test	278
10.15Significance Analysis-Anova	279
10.16Fold Change	280
10.17GO Analysis	282
10.18Fold Change	283
10.19Load Data	284
10.20Identify Calls Range	285
10.21Preprocess Options	286
10.22Choose Entities	287
10.23Preprocess Baseline Options	288
10.24Gene Level Experiment Creation	289

10.25	Gene Level Experiment Creation - Normalization Options	290
10.26	Gene Level Experiment Creation - Choose Entities	291
10.27	Gene Level Experiment Creation - Preprocess Baseline Options	292
10.28	Quality Control	293
10.29	Entity list and Interpretation	294
10.30	Input Parameters	295
10.31	Output Views of Filter by Flags	296
10.32	Save Entity List	297
11.1	Welcome Screen	302
11.2	Create New project	302
11.3	Experiment Selection	303
11.4	Experiment Description	305
11.5	Load Data	305
11.6	Choose Samples	306
11.7	Reordering Samples	306
11.8	Confirmation Dialog Box	307
11.9	Agilent Single Colour - Handling on chip replicates: Example 1	308
11.10	Agilent Single Colour - Handling on chip replicates: Example 2	308
11.11	Summary Report	309
11.12	Experiment Grouping	311

11.13	Edit or Delete of Parameters	312
11.14	Quality Control on Samples	313
11.15	Filter Probesets-Single Parameter	315
11.16	Filter Probesets-Two Parameters	316
11.17	Rerun Filter	316
11.18	Significance Analysis-T Test	319
11.19	Significance Analysis-Anova	320
11.20	Fold Change	321
11.21	GO Analysis	322
11.22	Find Significant Pathways	323
11.23	Load Data	324
11.24	Advanced flag Import	325
11.25	Preprocess Options	326
11.26	Normalize to control genes	327
11.27	Baseline Transformation Options	328
11.28	Gene Level Experiment Creation	330
11.29	Gene Level Experiment Creation - Normalization Options	331
11.30	Gene Level Experiment Creation - Choose Entities	332
11.31	Gene Level Experiment Creation - Preprocess Baseline Options	333
11.32	Quality Control	334
11.33	Entity list and Interpretation	335

11.34	Input Parameters	335
11.35	Output Views of Filter by Flags	337
11.36	Save Entity List	338
12.1	Welcome Screen	342
12.2	Create New project	342
12.3	Experiment Selection	343
12.4	Experiment Description	345
12.5	Load Data	345
12.6	Choose Samples	346
12.7	Reordering Samples	346
12.8	Dye Swap	347
12.9	Agilent Two Colour - Handling on chip replicates: Example 1	348
12.10	Agilent Two Colour - Handling on chip replicates: Example 2	348
12.11	Summary Report	349
12.12	Experiment Grouping	351
12.13	Edit or Delete of Parameters	352
12.14	Quality Control on Samples	353
12.15	Filter Probesets-Single Parameter	355
12.16	Filter Probesets-Two Parameters	356
12.17	Rerun Filter	356

12.18	Significance Analysis-T Test	360
12.19	Significance Analysis-Anova	361
12.20	Fold Change	362
12.21	GO Analysis	363
12.22	Find Significant Pathways	364
12.23	Load Data	366
12.24	Samples Validation	367
12.25	Choose Dye-Swaps	368
12.26	Advanced flag Import	369
12.27	Preprocess Options	370
12.28	Gene Level Experiment Creation	371
12.29	Gene Level Experiment Creation - Normalization Options	372
12.30	Gene Level Experiment Creation - Choose Entities	373
12.31	Gene Level Experiment Creation - Preprocess Baseline Options	374
12.32	Quality Control	375
12.33	Entity list and Interpretation	376
12.34	Input Parameters	376
12.35	Output Views of Filter by Flags	378
12.36	Save Entity List	379
13.1	Welcome Screen	382

13.2 Create New project	382
13.3 Experiment Selection	383
13.4 Experiment Selection	385
13.5 Load Data	385
13.6 Technology Creation in miRNA	386
13.7 Selection of Organism	386
13.8 Confirmation Window	386
13.9 Summary Report	389
13.10Experiment Grouping	391
13.11Add/Edit Parameters	392
13.12Quality Control on Samples	393
13.13Filter Probesets-Single Parameter	395
13.14Filter Probesets-Two Parameters	396
13.15Significance Analysis-T Test	399
13.16Significance Analysis-Anova	400
13.17Fold Change	401
13.18TargetScan Database Download	401
13.19Biological Genome Download	401
13.20GO Analysis	403
13.21Find Significant Pathways	404
13.22Load Data	406

13.23	Normalization Options	407
13.24	Choose entities	408
13.25	Baseline Transformation	409
13.26	Selection of Controls	410
13.27	Quality Control	411
13.28	Entity list and Interpretation	413
13.29	Input Parameters	413
13.30	Output Views of Filter by Flags	414
13.31	Save Entity List	415
13.32	Workflow Navigator-TargetScan	418
13.33	Inputs for TargetScan	418
14.1	Experiment Creation	423
14.2	Baseline Transformation Options	424
14.3	Quality Control	427
14.4	Input Parameters	430
14.5	Choose Annotation Columns	432
15.1	Technology Name	434
15.2	Format data file	435
15.3	Select Row Scope for Import	436
15.4	Single Color one sample in one file selections	437

15.5 Single Color-Multiple Samples Per File-Keyword Selection	438
15.6 Single Color-Multiple Samples Per File-Custom Selection	439
15.7 Annotation Column Options	441
15.8 Annotation Mark Colors	442
15.9 Welcome Screen	443
15.10 Create New project	444
15.11 Experiment Selection	444
15.12 Experiment Description	445
15.13 Load Data	447
15.14 Preprocess Options	448
15.15 Choose Entities	449
15.16 Preprocess Baseline Options	450
15.17 Gene Level Experiment Creation	452
15.18 Gene Level Experiment Creation - Normalization Options	453
15.19 Gene Level Experiment Creation - Choose Entities	454
15.20 Gene Level Experiment Creation - Preprocess Baseline Options	455
15.21 Quality Control	456
15.22 Entity list and Interpretation	456
15.23 Input Parameters	457
15.24 Output Views of Filter by Flags	458
15.25 Save Entity List	459

16.1 Technology Name	462
16.2 Format data file	463
16.3 Select Row Scope for Import	464
16.4 Two Color Selections	465
16.5 Annotation Mark Colors	468
16.6 Annotation Column Options	469
16.7 Technology Creation	469
16.8 Welcome Screen	470
16.9 Create New project	471
16.10Experiment Selection	471
16.11Experiment Description	472
16.12Load Data	473
16.13Choose Dye-Swaps	474
16.14Preprocess Options	474
16.15Preprocess Baseline Options	475
16.16Gene Level Experiment Creation	478
16.17Gene Level Experiment Creation - Normalization Options	479
16.18Gene Level Experiment Creation - Choose Entities	480
16.19Gene Level Experiment Creation - Preprocess Baseline Options	481
16.20Quality Control	482
16.21Entity list and Interpretation	482

16.22	Input Parameters	483
16.23	Output Views of Filter by Flags	484
16.24	Save Entity List	485
17.1	GEO Identifier Entry Dialog	490
17.2	Create New Experiment Dialog	491
17.3	Experiment Grouping Information is automatically copied over	492
17.4	Duplicate Experiment Parameters	493
17.5	Duplicate Parameters	493
17.6	Final Experiment Grouping	494
17.7	Sample attributes that can be chosen as Experiment Parameters	496
18.1	Experiment Grouping	498
18.2	Edit or Delete of Parameters	500
18.3	Create Interpretation (Step 1 of 3)	500
18.4	Create Interpretation (Step 2 of 3)	501
18.5	Create Interpretation (Step 2 of 3)	502
18.6	Gene Level Experiment Creation	503
18.7	Gene Level Experiment Creation - Normalization Options	504
18.8	Gene Level Experiment Creation - Choose Entities	505
18.9	Gene Level Experiment Creation - Preprocess Baseline Options	506
18.10	Filter probesets by expression (Step 1 of 4)	507

18.11Filter probesets by expression (Step 2 of 4)	508
18.12Filter probesets by expression (Step 3 of 4)	509
18.13Filter probesets by expression (Step 4 of 4)	510
18.14Input Parameters	513
18.15Select Test	514
18.16p-value Computation	515
18.17Results	517
18.18Save Entity List	518
18.19Pairing Options	520
18.20Input Parameters	524
18.21Pairing Options	525
18.22Fold Change Results	526
18.23Object Details	527
18.24Input Parameters	529
18.25Output View of Find Similar Entities	530
18.26Save Entity List	531
18.27Input Parameters	532
18.28Output View of Filter on Parameters	533
18.29Save Entity List	534
18.30Entity List and Interpretation	535
18.31Input Parameters	536

18.32	Output Views	537
18.33	Filter on Entity List - Step 1	542
18.34	Filter on Entity List - Step 2	543
18.35	Filter on Entity List - Step 3	543
18.36	Filter on Entity List - Step 4	544
19.1	Anova result showing 'Excluded Entities' because of missing values	556
20.1	Clustering Wizard: Input parameters	565
20.2	Clustering Wizard: Clustering parameters	565
20.3	Clustering Wizard: Output Views	566
20.4	Clustering Wizard: Object details	567
20.5	Cluster Set from K-Means Clustering Algorithm	568
20.6	Dendrogram View of Clustering	571
20.7	Export Image Dialog	573
20.8	Error Dialog on Image Export	574
20.9	Dendrogram Toolbar	574
20.10	U Matrix for SOM Clustering Algorithm	577
21.1	Classification Pipeline	587
21.2	Build Prediction Model: Input parameters	589
21.3	Build Prediction Model: Validation parameters	590
21.4	Build Prediction Model: Validation output	591

21.5 Build Prediction Model: Training output	592
21.6 Build Prediction Model: Model Object	593
21.7 Run Prediction: Prediction output	594
21.8 Axis Parallel Decision Tree Model	596
21.9 Neural Network Model	598
21.10 Model Parameters for Support Vector Machines	601
21.11 Model Parameters for Naive Bayesian Model	603
21.12 Confusion Matrix for Training with Decision Tree	606
21.13 Decision Tree Classification Report	607
21.14 Lorenz Curve for Neural Network Training	608
22.1 Input Parameters	612
22.2 Output Views of GO Analysis	613
22.3 Spreadsheet view of GO Terms.	615
22.4 The GO Tree View.	616
22.5 Properties of GO Tree View.	617
22.6 Pie Chart View.	617
22.7 Pie Chart Properties.	619
23.1 Input Parameters	625
23.2 Pairing Options	626
23.3 Choose Gene Lists	627

23.4 Results	628
24.1 Input Parameters	633
24.2 Pairing Options	634
24.3 Choose Gene Sets	635
24.4 Choose Gene Sets	636
24.5 Choose Gene Lists	637
25.1 Simple Analysis	644
25.2 Advanced Analysis	645
25.3 Error Message	646
25.4 Matching Statistics	647
25.5 Analysis Filters-Direct Algorithm	648
25.6 Analysis Filters-Expand Algorithm	649
25.7 Analysis Filters-Shortest Connect	650
25.8 Analysis Result	651
25.9 Save Pathway	652
25.10Node-Legend	653
25.11Node Properties	654
25.12Edges-Legend	655
25.13Relations-Legend	655
25.14Relation Properties	656

25.15	Toolbar	656
25.16	Data Overlay Properties	660
25.17	Data Overlay	661
25.18	Legend for Data Overlay	662
25.19	Main menu-Search	662
25.20	Input Parameters	663
25.21	Output Views	663
25.22	Entity Inspector	664
25.23	Search Parameters	665
25.24	Advanced Search Parameters	665
25.25	Search Results	666
25.26	Twopi layout	667
25.27	Tools—>Edit Pathway Theme	668
25.28	Style Theme Dialog	669
25.29	Extract Interactions via NLP	669
25.30	Input Data	670
25.31	View Tagged Content	671
25.32	Pathway View	672
25.33	Object Details	673
25.34	Step 1: Input Page	674
25.35	Step 2: Select Relevant MeSH Terms	675

25.36	Step 3: MeSH Pathway	676
25.37	Select Pathways to Import	677
25.38	Choose BioPAX files	678
25.39	Select Pathways to Import	680
25.40	Input Parameters	681
25.41	Results Window	682
25.42	Pathway Experiment Creation	684
25.43	Input Parameters	684
25.44	Import List from File	685
25.45	Choose signal columns	686
25.46	Choose extra column	687
25.47	Pathway Experiment Workflow	688
25.48	Data Export	691
25.49	Update Pathway Interactions	704
26.1	Experiment Creation for Affy CEL files	711
26.2	Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set - Reference Creation	713
26.3	Reference Creation for Affy 100K array set	714
26.4	Create Technology for Copy Number Analysis - Affymetrix technology	717
26.5	Step 1: Load Data	719
26.6	Step 2: Pair CEL files	719

26.7 Step 3: Choose Copy Number/LOH Analysis Type	720
26.8 QC views for Copy Number Experiment	722
26.9 Batch Effect Correction - Step 1	725
26.10Batch Effect Correction - Step 2	726
26.11Copy Number Analysis - Paired Normal Method	728
26.12Copy Number Analysis - Against Reference Method	729
26.13Common Genomic Variant Regions - Step 3	732
26.14Common Genomic Variant Regions - Step 4	733
26.15Step 2: Filter Conditions for Filter by Region	736
26.16Step 2: Input parameters for PSCN Filter	738
26.17Heap Map View for a Copy Number Experiment	740
26.18BRLMM-Flow chart	751
26.19Transition Probabilities for LOH analysis against Reference HMM	753
26.20The Paired Normal HMM	753
27.1 EIGENSTRAT Filter	770
27.2 EIGENSTRAT Correction View	779
27.3 EIGENSTRAT Correction Results	781
27.4 LD Plot	798
27.5 Allele Frequencies on Genome Browser	799
28.1 Genome Browser showing the panels	806

28.2 Genome Browser - Select Build	807
28.3 Genome Browser - On Launch	808
28.4 Genome Browser - Import and Manage Tracks	810
28.5 Genome Browser - Add/Delete Organism	811
28.6 Genome Browser - Add New Build	812
28.7 Genome Browser - Step 1 of Advanced Import	813
28.8 Genome Browser - Step 2 of Advanced Import	814
28.9 Genome Browser - Step 3 of Advanced Import	815
28.10Genome Browser - Select Data	816
29.1 Launch IPA	824
29.2 Create Pathway in IPA	824
29.3 Create New Pathway	826
29.4 IPA Pathway Creation	827
29.5 Java Startup	827
29.6 IPA Login Dialog	828
29.7 Pathway Analysis in IPA	828
29.8 Creation of Entity List	829
29.9 Creation of Entity List	830
29.10Save List	831
29.11GeneSpring GX suitable list creation	831

29.12	Saved List Location	832
29.13	Import IPA Entity List	833
29.14	Selection of Folder	834
29.15	Entity List Creation	834
29.16	Error Message	835
29.17	Launch IPA	835
29.18	Data Analysis on Experiment	836
29.19	Perform Data Analysis on Experiment	836
29.20	IPA Pathway Creation	838
29.21	Java Startup	838
29.22	IPA Login Dialog	839
29.23	Create Analysis	840
29.24	Analysis Settings	841
29.25	Launch IPA	842
29.26	Data Analysis on Entity List	842
29.27	Perform Data Analysis on Entity List	844
29.28	IPA Pathway Creation	845
29.29	Java Startup	845
29.30	IPA Login Dialog	846
29.31	Create Analysis	847
29.32	Analysis Settings	848

30.1 Permission Dialog	853
30.2 Task Manager	856

List of Tables

1.1	Platform Compatibility	1
1.2	Windows Platform Compatibility	3
1.3	Linux Platform Compatibility	6
1.4	Mac OS X Platform Compatibility	9
2.1	Interpretations and Views	25
2.2	Interpretations and Workflow Operations	26
3.1	HomoloGene Table	54
4.1	Migration Rate	63
4.2	Migration Rate on Windows OS	64
4.3	Migration Rate on Debian OS	64
6.1	Sample Grouping and Significance Tests I	151
6.2	Sample Grouping and Significance Tests II	153
6.3	Sample Grouping and Significance Tests III	153
6.4	Sample Grouping and Significance Tests IV	153

6.5	Sample Grouping and Significance Tests V	154
6.6	Sample Grouping and Significance Tests VI	154
6.7	Sample Grouping and Significance Tests VII	154
6.8	Table of Default parameters for Guided Workflow	160
8.1	Sample Grouping and Significance Tests I	202
8.2	Sample Grouping and Significance Tests II	204
8.3	Sample Grouping and Significance Tests III	205
8.4	Sample Grouping and Significance Tests IV	205
8.5	Sample Grouping and Significance Tests V	206
8.6	Sample Grouping and Significance Tests VI	206
8.7	Sample Grouping and Significance Tests VII	207
8.8	Table of Default parameters for Guided Workflow	211
10.1	Sample Grouping and Significance Tests I	276
10.2	Sample Grouping and Significance Tests II	276
10.3	Sample Grouping and Significance Tests III	276
10.4	Sample Grouping and Significance Tests IV	277
10.5	Sample Grouping and Significance Tests V	277
10.6	Sample Grouping and Significance Tests VI	277
10.7	Sample Grouping and Significance Tests VII	278
10.8	Table of Default parameters for Guided Workflow	284

11.1 Quality Controls Metrics	314
11.2 Sample Grouping and Significance Tests I	314
11.3 Sample Grouping and Significance Tests II	315
11.4 Sample Grouping and Significance Tests III	317
11.5 Sample Grouping and Significance Tests IV	317
11.6 Sample Grouping and Significance Tests V	318
11.7 Sample Grouping and Significance Tests VI	318
11.8 Sample Grouping and Significance Tests VII	319
11.9 Table of Default parameters for Guided Workflow	323
11.10Quality Controls Metrics	336
12.1 Quality Controls Metrics	354
12.2 Sample Grouping and Significance Tests I	355
12.3 Sample Grouping and Significance Tests II	357
12.4 Sample Grouping and Significance Tests III	357
12.5 Sample Grouping and Significance Tests IV	358
12.6 Sample Grouping and Significance Tests V	358
12.7 Sample Grouping and Significance Tests VI	359
12.8 Sample Grouping and Significance Tests VII	359
12.9 Table of Default parameters for Guided Workflow	365
12.10Quality Controls Metrics	377

13.1 Sample Grouping and Significance Tests I	394
13.2 Sample Grouping and Significance Tests II	395
13.3 Sample Grouping and Significance Tests III	396
13.4 Sample Grouping and Significance Tests IV	397
13.5 Sample Grouping and Significance Tests V	397
13.6 Sample Grouping and Significance Tests VI	398
13.7 Sample Grouping and Significance Tests VII	398
13.8 Table of Default parameters for Guided Workflow	405
18.1 Sample Grouping and Significance Tests I	513
18.2 Sample Grouping and Significance Tests I	519
18.3 Sample Grouping and Significance Tests II	519
18.4 Sample Grouping and Significance Tests III	519
18.5 Sample Grouping and Significance Tests IV	520
18.6 Sample Grouping and Significance Tests V	521
18.7 Sample Grouping and Significance Tests VI	521
18.8 Sample Grouping and Significance Tests VII	521
18.9 Sample Grouping and Significance Tests VIII	522
21.1 Decision Tree Table	591
21.2 Validation Parameters	605
25.1 Right-Click Legend	658

25.2	Type of relationship and Frequency	675
25.3	Process, Function and Complex Properties	696
25.4	Participant Roles	698
25.5	Protein Entities in Pathway Database	700
25.6	Protein Entities in Pathway Database	700
25.7	Other Entities in Pathway database	700
25.8	Total Number of Relation classified as “Generic”	700
25.9	Total Number of Relation in Pathway database	701
25.10	Total Number of Relation in Pathway database	701
25.11	Relations from each Data source	701
25.12	Relations from each Data source	701
26.1	Terminology in Copy Number Analysis	709
26.2	Mapping Fawkes state to LOH	716
26.3	Batch Effect Correction	724
26.4	snapshot of 'Common Genomic Variant Region' Detection Algorithm	730
26.5	Filter by CGVs	739
26.6	Identify Overlapping Genes	742
26.7	Utilities in Copy Number Analysis	743
26.8	Workflow for Illumina output files	746
26.9	Additional notes on BRLMM	750

26.10	Snap-shot of Birdseed Algorithm	755
26.11	Snap-shot of CBS Algorithm	757
26.12	Snap-shot of Fawkes Algorithm	759
27.1	Technologies and Genotype Call Algorithms for Association Analysis Experiments	762
27.2	Summary of Steps: Filter Samples by Missing Values	766
27.3	Birdseed Report	767
27.4	Summary of Steps: EIGENSTRAT Filter	769
27.5	Summary of Steps: Filter SNPs by Missing Value	773
27.6	Summary of Steps: Filter SNPs by Differential Missingness	774
27.7	Contingency Table for Differential Missingness	774
27.8	Summary of Steps: Filter SNPs by HWE p-value	775
27.9	Summary of Steps: Filter SNPs by MAF	777
27.10	Summary of Steps: EIGENSTRAT Correction on Samples	778
27.11	EIGENSTRAT Correction Result Screen	780
27.12	Mode of Inheritance	782
27.13	Summary of Steps: Pearson's χ^2 Test	786
27.14	Contingency Table for Pearson's χ^2 Test	786
27.15	Summary of Steps: Fisher's Exact Test	787
27.16	Summary of Steps: Cochran-Armitage Test	788
27.17	Contingency Table for Cochran-Armitage Test	788

27.18	Weights (d_i) for Cochran-Armitage Test	788
27.19	Summary of Steps: χ^2 Correlation Test	789
27.20	Mapping for Genotype Calls	800
27.21	Utilities in Association Analysis	801
28.1	Annotation Track Properties	819
31.1	Accessing Projects and Experiments	861
31.2	Accessing Experiment Dataset	864
31.3	Some Useful Functions	867
31.4	Some Common Marks	867
31.5	Creating UI Components	869
32.1	Mouse Clicks and their Action	885
32.2	Scatter Plot Mouse Clicks	886
32.3	3D Mouse Clicks	886
32.4	Mouse Click Mappings for Mac	886
32.5	Global Key Bindings	887

Chapter 1

GeneSpring GX Installation

This version of **GeneSpring GX** 11.0 is available for Windows, Mac OS X (IntelMac), and Linux. This chapter describes how to install **GeneSpring GX** on Windows, Mac OS X and Linux. Note that this version of **GeneSpring GX** can coexist with **GeneSpring GX** 7.x on the same machine.

1.1 Supported and Tested Platforms

The table below gives the platforms on which **GeneSpring GX** has been tested.

Operating System	Hardware Architecture	Installer
Microsoft Windows XP Service Pack 3	x86 compatible architecture	genespringGX_windows32.exe
Microsoft Windows XP Service Pack 3	x86_64 compatible architecture	genespringGX_windows64.exe
Microsoft Windows Vista	x86 compatible architecture	genespringGX_windows32.exe
Microsoft Windows Vista	x86_64 compatible architecture	genespringGX_windows64.exe
Red Hat Enterprise Linux 5	x86 compatible architecture	genespringGX_linux32.bin
Red Hat Enterprise Linux 5	x86_64 compatible architecture	genespringGX_linux64.bin
Debian GNU/Linux 4.0r1	x86 compatible architecture	genespringGX_linux32.bin
Debian GNU/Linux 4.0r1	x86_64 compatible architecture	genespringGX_linux64.bin
Apple Mac OS X v10.4	x86 compatible architecture	genespringGX_mac.zip
Apple Mac OS X v10.6 (Snow Leopard)	x86 compatible architecture	genespringGX_mac.zip

Table 1.1: Platform Compatibility

1.1.1 System Requirements for Copy Number and Association Experiments

Supported Platforms

Copy Number and Association experiments runs on all supported platforms as mentioned above with the following exception:

- For MAC users, Copy Number and Association Experiments do not run on 10.4.x Tiger. Among MAC platforms, the recommended one for Copy Number and Association experiments is Apple Mac OS X v10.6 (Snow Leopard).

Minimum Specifications

As a guideline, the minimum specifications for, say, 75 samples of Affymetrix Genome-Wide Human SNP Array 6.0 are:

- 32-bit system with 2 GB RAM
- 25 GB of free disk space

Recommended Specifications

- A 64-bit, Quadcore platform with 4 GB or higher RAM is recommended.
- Free disk space required will be proportional to the number and size of samples. An approximation can be made based on the guidelines provided in the 'Minimum Specifications' section above.

1.1.2 Installation and Usage Requirements

Supported Windows Platforms

- Operating System: Microsoft Windows XP Service Pack 2, Microsoft Windows Vista, 32-bit and 64-bit operating systems.
- Pentium 4 with 1.5 GHz and 1 GB RAM.

Operating System	Hardware Architecture	Installer
Microsoft Windows XP Service Pack 3	x86 compatible architecture	genespringGX_windows32.exe
Microsoft Windows XP Service Pack 3	x86_64 compatible architecture	genespringGX_windows64.exe
Microsoft Windows Vista	x86 compatible architecture	genespringGX_windows32.exe
Microsoft Windows Vista	x86_64 compatible architecture	genespringGX_windows64.exe

Table 1.2: Windows Platform Compatibility

- Disk space required: 1 GB
- At least 16MB Video Memory. Check this via Start →Settings →Control Panel →Display →Settings tab →Advanced →Adapter tab →Memory Size field. 3D graphics may require more memory. Also changing Display Acceleration settings may be needed to view 3D plots.
- Administrator privileges are required for installation. Once installed, other users can use **GeneSpring GX** as well.

1.1.3 GeneSpring GX Installation Procedure for Microsoft Windows

GeneSpring GX can be installed on any of the Microsoft Windows platforms listed above. To install **GeneSpring GX**, follow the instructions given below:

- You must have the installable for your particular platform `genespringGX_windows.exe`.
- Run the `genespringGX_windows.exe` installable file.
- The wizard will guide you through the installation procedure.
- By default, **GeneSpring GX** will be installed in the `C:\Program Files\Agilent\GeneSpringGX\` directory. You can specify any other installation directory of your choice during the installation process.
- At the end of the installation process, a browser is launched with the documentation index, showing all the documentation available with the tool.
- Following this, **GeneSpring GX** is installed on your system. By default the **GeneSpring GX** icon appears on your desktop and in the programs menu.
- To start using **GeneSpring GX**, you will have to activate your installation by following the steps detailed in the [Activation](#) step.

By default, **GeneSpring GX** is installed in the programs group with the following utilities:

- **GeneSpring GX**, for starting up the **GeneSpring GX** tool.
- Documentation, leading to all the documentation available online in the tool.
- Uninstall, for uninstalling the tool from the system.

1.1.4 Activating your GeneSpring GX

Your **GeneSpring GX** installation has to be activated for you to use **GeneSpring GX**. **GeneSpring GX** imposes a node-locked license, so it can be used only on the machine that it was installed on.

- You should have a valid OrderID to activate **GeneSpring GX**. If you do not have an OrderID, register at <http://genespring.com>. An OrderID will be e-mailed to you to activate your installation.
- Auto-activate **GeneSpring GX** by connecting to **GeneSpring GX** website. The first time you start up **GeneSpring GX** you will be prompted with the ‘**GeneSpring GX** License Activation’ dialog-box. Enter your OrderID in the space provided. This will connect to the **GeneSpring GX** website, activate your installation and launch the tool. If you are behind a proxy server, then provide the proxy details in the lower half of this dialog-box.
- The license is obtained by contacting the licenses server over the Internet and obtaining a node-locked, fixed duration license. If your machine date and time settings are different and cannot be matched with the server date and time settings you will get an *Clock Skew Detected* error and will not be able to proceed. If this is a new installation, you can change the date and time on your local machine and try to activate again.
- **Manual activation.** If the auto-activation step has failed due to any other reason, you will have to manually get the activation license file to activate **GeneSpring GX**, using the instructions given below:
 - Locate the activation key file `manualActivation.txt` in the `\bin\license\` folder in the installation directory.
 - Go to <http://lcosgens.cos.agilent.com/gsLicense/Activate.html>, enter the OrderID, upload the activation key file, `manualActivation.txt` from the file-path mentioned above, and click Submit. This will generate an activation license file (`strand.lic`) that will be e-mailed to your registered e-mail address. If you are unable to access the website or have not received the activation license file, send a mail to informatics_support@agilent.com with the subject **Registration Request**, with `manualActivation.txt` as an attachment. We will generate an activation license file and send it to you within one business day.
 - Once you have got the activation license file, `strand.lic`, copy the file to your `\bin\license\` subfolder.
 - Restart **GeneSpring GX**. This will activate your **GeneSpring GX** installation and will launch **GeneSpring GX**.

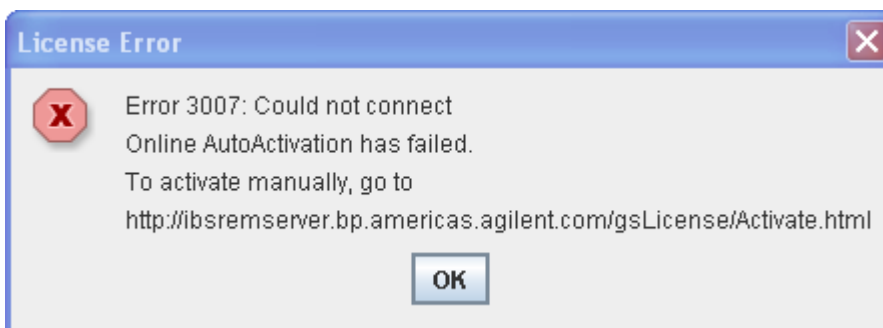


Figure 1.1: Activation Failure

- If **GeneSpring GX** fails to launch and produces an error, please send the error code to informatics_support@agilent.com with the subject Activation Failure. You should receive a response within one business day.

1.1.5 Uninstalling GeneSpring GX from Windows

The Uninstall program is used for uninstalling **GeneSpring GX** from the system. Before uninstalling **GeneSpring GX**, make sure that the application and any open files from the installation directory are closed.

To start the **GeneSpring GX** uninstaller, click Start, choose the Programs option, and select **GeneSpringGX**. Click Uninstall. Alternatively, click Start, select the Settings option, and click Control Panel. Double-click the Add/Remove Programs option. Select **GeneSpringGX** from the list of products. Click Uninstall. The Uninstall **GeneSpring GX** wizard displays the features that are to be removed. Click Done to close the Uninstall Complete wizard. **GeneSpring GX** will be successfully uninstalled from the Windows system. Some files and folders like log files and data, samples and templates folders that have been created after the installation of **GeneSpring GX** would not be removed.

1.2 Installation on Linux

Supported Linux Platforms

Operating System	Hardware Architecture	Installer
Red Hat Enterprise linux 5	x86 compatible architecture	genespringGX_linux32.bin
Red Hat Enterprise linux	x86_64 compatible architecture	genespringGX_linux64.bin
Debian GNU/Linux 4.0r1	x86 compatible architecture	genespringGX_linux32.bin
Debian GNU/Linux 4.0r1	x86_64 compatible architecture	genespringGX_linux64.bin

Table 1.3: Linux Platform Compatibility

1.2.1 Installation and Usage Requirements

- RedHat Enterprise Linux 5.x. 32-bit as well as 64-bit architecture are supported.
- In addition certain run-time libraries are required for activating and running **GeneSpring GX**. The required run-time libraries are `libstdc++.so.6`. To confirm that the required libraries are available for activating the license, go to `Agilent/GeneSpringGX/bin/packages/cube/license/x.x/lib(32/64)` and run the following command

```
ldd liblicense.so
```

Check that all required linked libraries are available on the system.
- Pentium 4 with 1.5 GHz and 1 GB RAM.
- Disk space required: 1 GB
- At least 16MB Video Memory.
- Administrator privileges are NOT required. Only the user who has installed **GeneSpring GX** can run it. Multiple installs with different user names are permitted.

1.2.2 GeneSpring GX Installation Procedure for Linux

GeneSpring GX can be installed on most distributions of Linux. To install **GeneSpring GX**, follow the instructions given below:

- You must have the installable for your particular platform `genespringGX_linux.bin` or `genespringGX_linux.sh`.
- Run the `genespringGX_linux.bin` or `genespringGX_linux.sh` installable.
- The program will guide you through the installation procedure.
- By default, **GeneSpring GX** will be installed in the `$HOME/Agilent/GeneSpringGX` directory. You can specify any other installation directory of your choice at the specified prompt in the dialog box.
- At the end of the installation process, a browser is launched with the documentation index, showing all the documentation available with the tool.

- **GeneSpring GX** should be installed as a normal user and only that user will be able to launch the application.
- Following this, **GeneSpring GX** is installed in the specified directory on your system. However, it will not be active yet. To start using **GeneSpring GX** , you will have to activate your installation by following the steps detailed in the [Activation](#) step.

By default, **GeneSpring GX** is installed with the following utilities in the **GeneSpring GX** directory:

- **GeneSpring GX**, for starting up the **GeneSpring GX** tool.
- Documentation, leading to all the documentation available online in the tool.
- Uninstall, for uninstalling the tool from the system

1.2.3 Activating your GeneSpring GX

Your **GeneSpring GX** installation has to be activated for you to use **GeneSpring GX**. **GeneSpring GX** imposes a node-locked license, so it can be used only on the machine that it was installed on.

- You should have a valid OrderID to activate **GeneSpring GX**. If you do not have an OrderID, register at <http://genespring.com>. An OrderID will be e-mailed to you to activate your installation.
- Auto-activate **GeneSpring GX** by connecting to **GeneSpring GX** website. The first time you start up **GeneSpring GX** you will be prompted with the ‘**GeneSpring GX** License Activation’ dialog-box. Enter your OrderID in the space provided. This will connect to the **GeneSpring GX** website, activate your installation and launch the tool. If you are behind a proxy server, then provide the proxy details in the lower half of this dialog-box.
- The license is obtained by contacting the licenses server over the Internet and obtaining a node-locked, fixed duration license. If your machine date and time settings are different and cannot be matched with the server date and time settings you will get an *Clock Skew Detected* error and will not be able to proceed. If this is a new installation, you can change the date and time on your local machine and try activate again.
- **Manual activation.** If the auto-activation step has failed due to any other reason, you will have to manually get the activation license file to activate **GeneSpring GX**, using the instructions given below:
 - Locate the activation key file `manualActivation.txt` in the `\bin\license\` folder in the installation directory.
 - Go to <http://lcosgens.cos.agilent.com/gsLicense/Activate.html>, enter the OrderID, upload the activation key file,

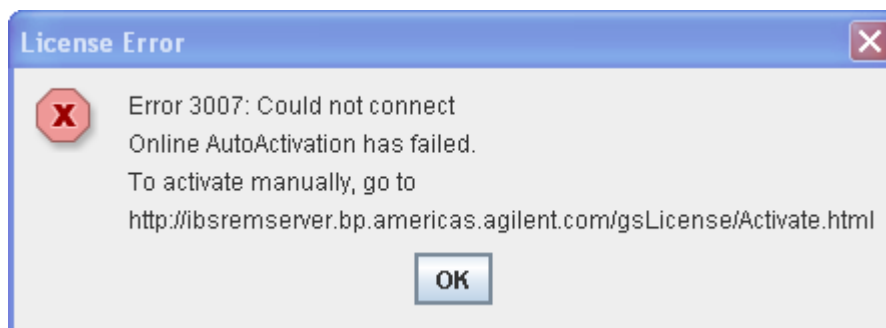


Figure 1.2: Activation Failure

`manualActivation.txt` from the file-path mentioned above, and click Submit. This will generate an activation license file (`strand.lic`) that will be e-mailed to your registered e-mail address. If you are unable to access the website or have not received the activation license file, send a mail to `informatics_support@agilent.com` with the subject **Registration Request**, with `manualActivation.txt` as an attachment. We will generate an activation license file and send it to you within one business day.

- Once you have got the activation license file, `strand.lic`, copy the file to your `\bin\license\` subfolder.
- Restart **GeneSpring GX**. This will activate your **GeneSpring GX** installation and will launch **GeneSpring GX**.
- If **GeneSpring GX** fails to launch and produces an error, please send the error code to `informatics_support@agilent.com` with the subject **Activation Failure**. You should receive a response within one business day.

1.2.4 Uninstalling GeneSpring GX from Linux

Before uninstalling **GeneSpring GX**, make sure that the application is closed. To uninstall **GeneSpring GX**, run Uninstall from the **GeneSpring GX** home directory and follow the instructions on screen.

1.3 Installation on Apple Macintosh

Supported Mac Platforms

Operating System	Hardware Architecture	Installer
Apple Mac OS X v10.4	x86 compatible architecture	genespringGX_mac.zip
Apple Mac OS X v10.6 (Snow Leopard)	x86 compatible architecture	genespringGX_mac.zip

Table 1.4: Mac OS X Platform Compatibility

1.3.1 Installation and Usage Requirements

- Mac OS X 10.4 and 10.6 (Snow Leopard); 10.5 is not supported. On Mac OS X 10.5 (Leopard), after running some features which use native code the program may get into an error state where running many other functions results in error. When this happens, the following error message is produced: "Create Native Shared Object". If this happens, re-install the application. This is a bug in Leopard and we have asked Apple for a solution to this issue.
- Processor with 1.5 GHz and 1 GB RAM.
- Disk space required: 1 GB
- At least 16MB Video Memory. (Refer section on 3D graphics in FAQ)
- Java version 1.5.0.05 or later; Check using "java -version" on a terminal, if necessary update to the latest JDK by going to Applications → System Prefs → Software Updates (system group).
- **GeneSpring GX** should be installed as a normal user and only that user will be able to launch the application.

1.3.2 GeneSpring GX Installation Procedure for Macintosh

- You must have the installable for your particular platform `genespringGX_mac.zip`.
- **GeneSpring GX** should be installed as a normal user and only that user will be able to launch the application.
- Uncompress the executable by double clicking on the .zip file. This will create a .app file at the same location. Make sure this file has executable permission.
- Double click on the .app file and start the installation. This will install **GeneSpring GX** on your machine. By default **GeneSpring GX** will be installed in `$HOME/Applications/Agilent/GeneSpringGX` or
You can install **GeneSpring GX** in an alternative location by changing the installation directory.
- To start using **GeneSpring GX**, you will have to activate your installation by following the steps detailed in the [Activation](#) step.
- At the end of the installation process, a browser is launched with the documentation index, showing all the documentation available with the tool.

- Note that **GeneSpring GX** is distributed as a node locked license. For this the hostname of the machine should not be changed. If you are using a DHCP server while being connected to be net, you have to set a fixed hostname. To do this, give the command `hostname` at the command prompt during the time of installation. This will return a hostname. And set the `HOSTNAME` in the file `/etc/hostconfig` to `your_machine_hostname_during_installation`

For editing this file you should have administrative privileges. Give the following command:

```
sudo vi /etc/hostconfig
```

This will ask for a password. You should give your password and you should change the following line

from

```
HOSTNAME=-AUTOMATIC-
```

to

```
HOSTNAME=your_machine_hostname_during_installation
```

- You need to restart the machine for the changes to take effect.

By default, **GeneSpring GX** is installed with the following utilities in the **GeneSpring GX** directory:

- **GeneSpring GX**, for starting up the **GeneSpring GX** tool.
- Documentation, leading to all the documentation available online in the tool.
- Uninstall, for uninstalling the tool from the system

GeneSpring GX uses left, right and middle mouse-clicks. On a single button Macintosh mouse, here is how you can emulate these clicks.

- Left-click is a regular single button click.
- Right-click is emulated by Control + click.
- Control-click is emulated by Apple + click.

1.3.3 Activating your GeneSpring GX

Your **GeneSpring GX** installation has to be activated for you to use **GeneSpring GX**. **GeneSpring GX** imposes a node-locked license, so it can be used only on the machine that it was installed on.

- You should have a valid OrderID to activate **GeneSpring GX**. If you do not have an OrderID, register at <http://genespring.com> An OrderID will be e-mailed to you to activate your installation.

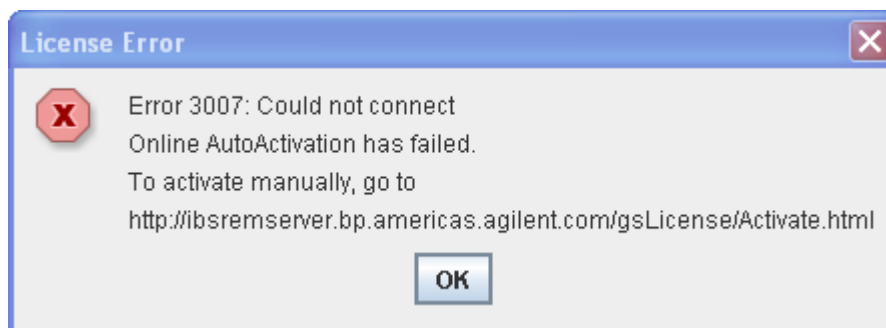


Figure 1.3: Activation Failure

- Auto-activate **GeneSpring GX** by connecting to **GeneSpring GX** website. The first time you start up **GeneSpring GX** you will be prompted with the 'GeneSpring GX License Activation' dialog-box. Enter your OrderID in the space provided. This will connect to the **GeneSpring GX** website, activate your installation and launch the tool. If you are behind a proxy server, then provide the proxy details in the lower half of this dialog-box.
- The license is obtained by contacting the licenses server over the internet and obtaining a node-locked, fixed duration license. If your machine date and time settings are different cannot be matched with the server date and time settings you will get an *Clock Skew Detected* error and will not be able to proceed. if this is a new installation, you can change the date and time on your local machine and try activate again.
- **Manual activation.** If the auto-activation step has failed due to any other reason, you will have to manually get the activation license file to activate **GeneSpring GX**, using the instructions given below:
 - Locate the activation key file `manualActivation.txt` in the `\bin\licence` subfolder of the installation directory.
 - Go to <http://lcosgens.cos.agilent.com/gsLicense/Activate.html>, enter the OrderID, upload the activation key file, `manualActivation.txt` from the file-path mentioned above, and click Submit. This will generate an activation license file (`strand.lic`) that will be e-mailed to your registered e-mail address. If you are unable to access the website or have not received the activation license file, send a mail to informatics.support@agilent.com with the subject **Registration Request**, with `manualActivation.txt` as an attachment. We will generate an activation license file and send it to you within one business day.
 - Once you have got the activation license file, `strand.lic`, copy the file to your `\bin\license\` subfolder of the installation directory.
 - Restart **GeneSpring GX**. This will activate your **GeneSpring GX** installation and will launch **GeneSpring GX**.
 - If **GeneSpring GX** fails to launch and produces an error, please send the error code to informatics.support@agilent.com with the subject **Activation Failure**. You should receive a response within one business day.

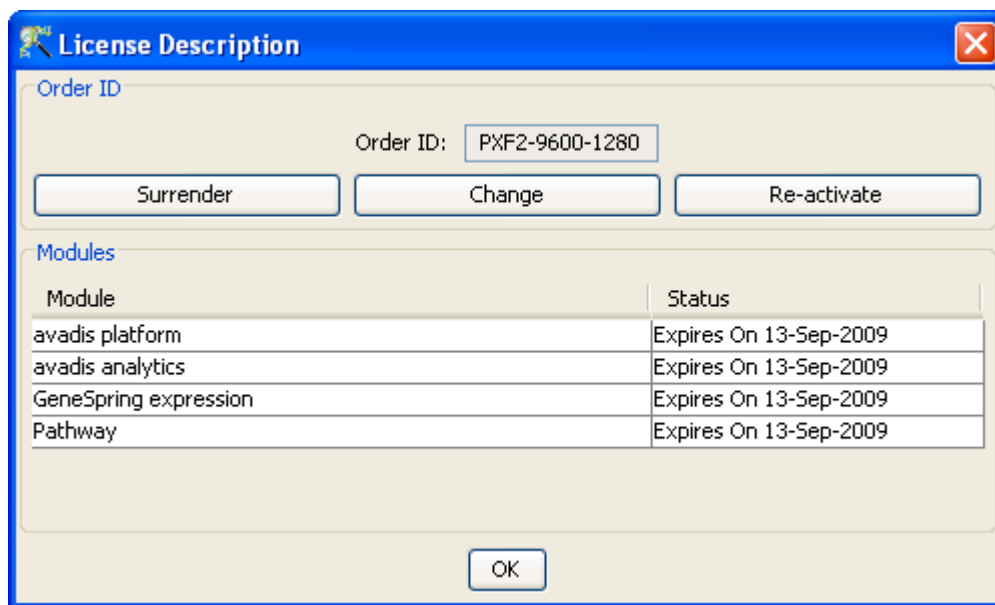


Figure 1.4: The License Description Dialog

1.3.4 Uninstalling GeneSpring GX from Mac

Before uninstalling **GeneSpring GX**, make sure that the application is closed. To uninstall **GeneSpring GX**, run Uninstall from the **GeneSpring GX** home directory and follow the instructions on screen.

1.4 License Manager

After successful installation and activation of **GeneSpring GX**, you will be able to use certain utilities to manage the license. These utilities are available from *Help* → *License Manager* on the top menu bar of the tool. Choosing *Help* → *License Manager* from the top menu will launch the *License Description* dialog.

The top box of the *License Manager* shows the **Order ID** that was used to activate the license. If you are using a floating server to activate and license **GeneSpring GX**, you will see the port and the host name of the license server. You may need to note the license **Order ID** to change the installation, or to refer to your installation at the time of support.

GeneSpring GX is licensed as a set of module bundles that allow various functionalities. The table in the dialog shows the modules available in the current installation along with their status. Currently the modules are bundled into the following categories:

- **avadis platform:** This provides the basic modules to launch the product and manage the user interfaces. This module is essential for the tool.
- **avadis analytics:** This module contains advanced analytics of clustering, classification and regression modules.
- **GeneSpring expression analysis:** This module enables the following gene expression analysis workflows:
 - Affymetrix® 3' IVT arrays,
 - Affymetrix Exon arrays for expression analysis,
 - Affymetrix Exon arrays for Splicing analysis,
 - Agilent single-color arrays,
 - Agilent two-color arrays,
 - Agilent miRNA arrays
 - Illumina® gene expression arrays,
 - Generic single-color arrays
 - Generic two-color arrays.
 - Copy Number Analysis
 - Association Analysis
- **Pathway:** This module enables the user to perform Pathway Analysis.

Based on the modules licensed, appropriate menu items will be enabled or disabled.

1.4.1 Utilities of the License Manager

The *License Manager* provides the following utilities. These are available from the *License Description* dialog.

Surrender: Click on this button to surrender the license to the license server. You must be connected to the internet for surrender to operate. The surrender utility is used if you want to check-in or surrender the license into the license server and check out or activate the license on another machine. This utility is useful to transfer licenses from one machine to another, like from an office desktop machine to a laptop machine.

Note that the license can be activated from only one installation at any time. Thus, when you surrender the license, the current installation will be in-activated. You will be prompted to confirm your intent to surrender the license and clicking *OK* will surrender the license and shut the tool. If you want to activate your license on another machine, or on the same machine, you will need to store the Order ID and enter the Order ID in the *License Activation Dialog*.

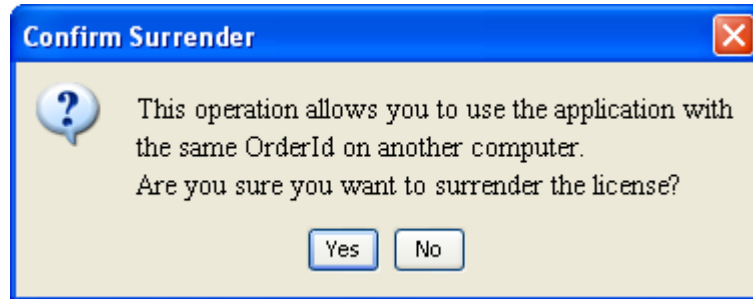


Figure 1.5: Confirm Surrender Dialog

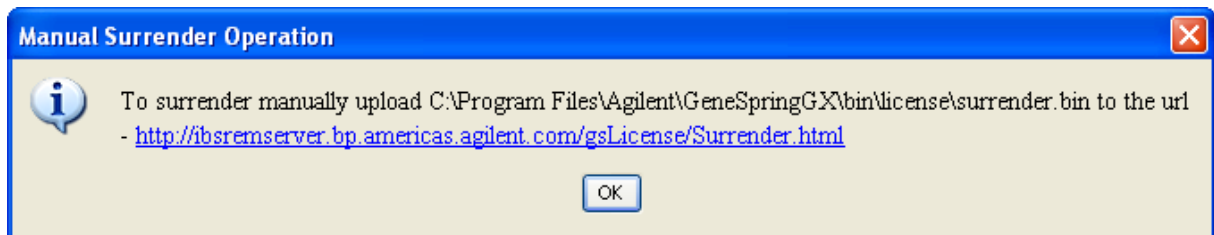


Figure 1.6: Manual Surrender Dialog

If you are not connected to the Internet, or if you are unable to reach the license server, you can do a manual surrender. You will be prompted with a dialog to confirm manual surrender. If you confirm, then the current installation will be deactivated. Follow the on screen instructions. Upload the file `<install_dir>/Agilent/GeneSpringGX/bin/license/surrender.bin` to <http://lcosgens.cos.agilent.com/gsLicense/Activate.html>. This will surrender the license which can be reused on another machine.

Change: This utility allows you to change the Order ID of the product and activate the product with a new Order ID. This utility is used to procure a different set of modules or change the module status and module expiry of the current installation. If you had a limited duration trial license and would like to purchase and convert the license to an annual license, click on the *Change* button. This will launch a dialog for Order ID. Enter the new Order ID obtained from Agilent. This will activate **GeneSpring GX** with the new Order ID and all the modules and module status will confirm to the new Order ID.

Re-activate: To reactivate the license, click on the *Re-activate* button on the **License Description Dialog**. This will reactivate the license from the license server with the same Order ID and on the same machine. The operation will prompt a dialog to confirm the action, after which the license will be reactivated and the tool will be shut down. When the tool is launched again, the tool will be launched again with the license obtained for the same Order ID. Note that reactivation can be done only on the same machine with the same Order ID. This utility may be necessary if the current installation is and license have been corrupted and you would like to reactivate and get a fresh license on the same Order ID on the same machine. Or you have Order ID definition and corresponding modules have changed and you have been advised by support to re-activate the license.

If you are not connected to the Internet, or if you are unable to reach the license server, you can re-activate manually. You will be prompted with a dialog stating that the reactivation failed and if



Figure 1.7: Change License Dialog

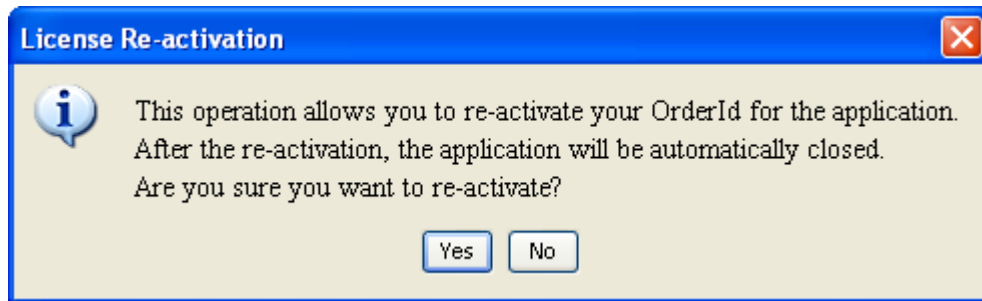


Figure 1.8: License Re-activation Dialog

you want to reactivate manually. If you confirm, then the current installation will be deactivated. Follow the on screen instructions to re-activate your tool.
<install_dir>/Agilent/GeneSpringGX/bin/license/surrender.bin
to <http://lcosgens.cos.agilent.com/gsLicense/Activate.html>.

1.5 Upgrade

For upgrading **GeneSpring GX** , go to the menu **Help** → **Update Product** → **From Agilent Server** for **From File** in the tool and follow the instructions thereon.

Chapter 2

GeneSpring GX Quick Tour

2.1 Introduction

This chapter gives a brief introduction to **GeneSpring GX**, explains the terminology used to refer to various organizational elements in the user interface, and provides a high-level overview of the data and analysis paradigms available in the application. The description here assumes that **GeneSpring GX** has already been installed and activated properly. To install and get **GeneSpring GX** activated, see [GeneSpring GX Installation](#).

2.2 Launching GeneSpring GX

To launch **GeneSpring GX**, you should have activated your license and your license must be valid. Launch the tool from the start menu or the desktop icon. On first launch **GeneSpring GX**, opens up with the demo project. On subsequent launches, the tool is initialized and shows a startup dialog. This dialog allows you to create a new project, open an existing project or open a recent project from the drop-down list. If you do not want the startup dialog uncheck the box on the dialog. You can restore the startup dialog by going to *Tools* → *Options* → *Miscellaneous* → *Startup Dialog*

2.3 GeneSpring GX User Interface

A screenshot of **GeneSpring GX** with various experiment and views is shown below. See Figure [2.1](#)

The main window consists of four parts - the Menubar, the Toolbar, the Display Pane and the [Status Line](#). The Display Pane contains several graphical views of the dataset, as well as algorithm results. The

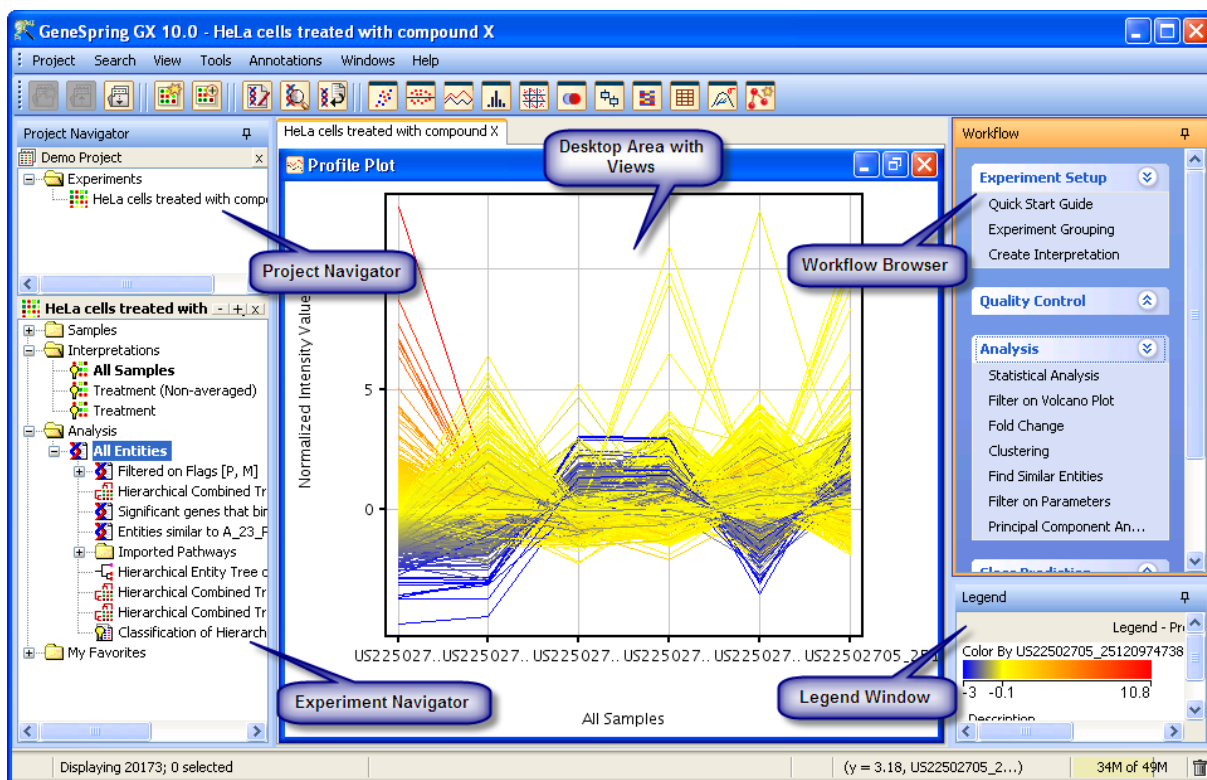


Figure 2.1: GeneSpring GX Layout

Display Pane is divided into three parts:

- The main **GeneSpring GX Desktop** in the center,
- The Project **Navigator** and the Experiment Navigator on the left,
- The **GeneSpring GX Workflow Browser**, and the **Legend Window** on the right.

2.3.1 GeneSpring GX Desktop

The desktop accommodates all the views pertaining to each experiment loaded in **GeneSpring GX**. Each window can be manipulated independently to control its size. Less important windows can be minimized. Windows can be tiled or cascaded in the desktop using the *Windows* menu. One of the views in the desktop is the active view.



Figure 2.2: The Workflow Window

2.3.2 Project Navigator

The Project Navigator on the left displays the project and all the experiments within it. Each experiment has its own navigator windows. The Project Navigator window shows all the experiments in the project. The experiment navigator window shows by default a **Samples** folder, an **Interpretation** folder, an **Analysis** folder and a **My Favorites** folder. The **My Favorites** folder can be populated with entity lists, Hierarchical trees, pathways or any other analysis objects that have been generated within the experiment., by copying (*Right-Click*→*Copy*) and pasting it (*Right-Click*→*Paste*), on to the appropriate sub-folder of *My Favorites*. New sub-folders can be created by going to *My Favorites*→*Right-Click*→*New Folder*

2.3.3 The Workflow Browser

The Workflow Browser shows the list of operations available in the experiment. It is organized into sequential groups of operations to help in the analysis of microarray data. The links in the Workflow

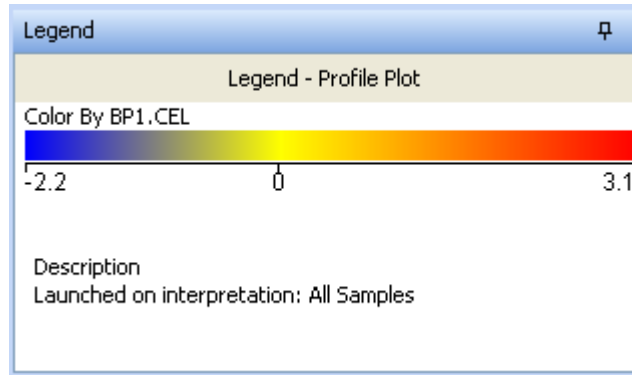


Figure 2.3: The Legend Window

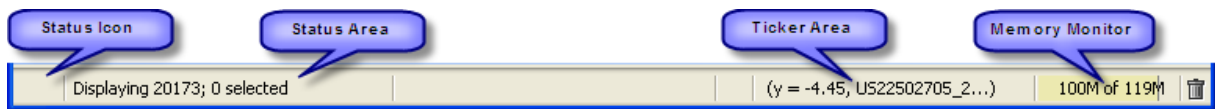


Figure 2.4: Status Line

Browser change according to the type of experiment being analyzed.

2.3.4 Global lists

'Global lists' enables users to tag entity lists as public and make it available across projects. This is different from the 'My favourites' utility under the experiment navigator, as Global lists appear across projects while 'My favourites' appear under each experiment in a particular project only.

See Section [Entity Lists](#) for details on operations possible on the 'Global lists'.

2.3.5 The Legend Window

The Legend window shows the legend for the current view in focus. Right-Click on the legend window shows options to *Copy* or *Export* the legend. Copying the legend will copy it to the Windows clipboard enabling pasting into any other Windows application using Control-V. Export will enable saving the legend as an image in one of the standard formats (JPG, PNG, JPEG etc).

2.3.6 Status Line

The status line is divided into four informative areas as depicted below. See Figure 2.4

Status Icon The status of the view is displayed here by an icon. Some views can be in the zoom or in the selection mode. The appropriate icon of the current mode of the view is displayed here.

Status Area This area displays high-level information about the current view. If a view is selection enabled, the status area shows the total number of rows or columns displayed and the number of entities / conditions selected. If the view is limited to selection, it will show that the view is limited to selection.

Ticker Area This area displays transient messages about the current view based upon the cursor location, eg., indicates the row and column indices for a spreadsheet or the X-Y co-ordinates of a scatter plot.

Memory Monitor This displays the total memory allocated to the Java process and the amount of memory currently used. You can clear memory running the Garbage Collector by Left-Click on the *Garbage Can* icon on the left. This will reduce the memory currently used by the tool.

2.4 Organizational Elements and Terminology in GeneSpring GX

Work in **GeneSpring GX** is organized into *projects*. A project comprises one or more related *experiments*. An experiment comprises *samples* (i.e., data sources), *interpretations* (i.e., groupings of samples based on experimental parameters), and *analyses* (i.e., statistical steps and associated results, typically entity lists). Statistical steps and methods of analysis are driven by a *workflow* which finds prominent mention on the right side of **GeneSpring GX**. These concepts are expanded below.

2.4.1 Project

A project is the key organizational element in **GeneSpring GX**. It is a container for a collection of experiments. For instance, researcher John might have a project on Lung Cancer. As part of this project, John might run several experiments. One experiment measures gene expression profiles of individuals with and without lung cancer, and one experiment measures the gene expression profiles of lung cancer patients treated with various new drug candidates. A single “Lung Cancer” project comprises both of these experiments. The ability to combine experiments into a project in **GeneSpring GX** allows for easy interrogation of “cross-experimental facts”, e.g., how do genes which are differentially expressed in individuals with lung cancer react to a particular drug.

A project can be created , viewed, deleted or closed using the following commands:

- **Project**→**New Project** creates a new project where the name and user notes can be specified.
- **Project**→**Open Project** opens an already created project.
- **Project**→**Recent Projects** allows access to recently opened projects.
- **Project**→**Close Project**

- **Project**→**Delete Project** deletes all the experiments and samples contained within the project.

Projects can also be exported out using **Project**→**Export Project Zip**. Likewise, projects can also be imported using **Import Project Zip** command.

A project could have multiple experiments that are run on different technology types, and possibly different organisms as well.

2.4.2 Experiment

An experiment in **GeneSpring GX** represents a collection of samples for which arrays have been run in order to answer a specific scientific question. A new experiment is created from **Project** →**New Experiment** by loading samples of a particular technology and performing a set of customary pre-processing steps like, normalization, summarization, baseline transform, etc., that will convert the raw data to a state where it is ready for analysis. An already created experiment can be opened and added to the open project from **Project** →**Add Experiment**.

A **GeneSpring GX** project could have many experiments. You can choose to selectively open/close each experiment. Each open experiment has its own section in the **Navigator**. **GeneSpring GX** allows exactly one of the open experiments to be *active* at any given point in time. The name of the active experiment is reflected in the title bar of the **GeneSpring GX** application. Also, the active experiment is highlighted with a broad orange line, letters in purple and a green icon.

An experiment consists of multiple samples, with which it was created, multiple interpretations, which group these samples by user-defined experimental parameters, and all other objects created as a result of various analysis steps in the experiment.

Datasets from GEO

Users can import datasets obtained from Gene Expression Omnibus (GEO) from <http://www.ncbi.nlm.nih.gov/geo/> into **GeneSpring GX**. Only expression datasets will be supported; exon, splicing data will not be imported. During the import, if **GeneSpring GX** detects that the technology of the datasets are not part of the standard technologies supported within the product, a message will be shown that "Unable to create experiment because matching Technology is not available. Do you want to import it as a Generic Experiment?" The user can then decide to create a generic experiment with such datasets.

Chapter [Loading Experiment from NCBI GEO](#) gives details on how to work with datasets from GEO.

2.4.3 Sample

An experiment comprises a collection of samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. A sample on the other hand, is created within **GeneSpring GX**, when it associates the data files with its appropriate technology. Thus, when an experiment is created with the raw hybridization data files, they get registered as samples of the appropriate technology in **GeneSpring GX**. Once registered, samples are available for use in other experiments as well. Thus an experiment can be created with new raw data files as well as samples already registered and available with **GeneSpring GX**.

2.4.4 Experiment Grouping, Parameters and Parameter Values

Samples in an experiment have associated experiment parameters and corresponding parameter values. For instance, if an experiment contains 6 samples, 3 treated with Drug X and 3 not treated, you would have one experimental parameter which you could call “Treatment Type”. Each sample needs to be given a value for this parameter. So you could call the 3 no treatment samples “Control” and the 3 treated samples “Drug X”. “Treatment Type” is the experimental parameter and “Control”/“Drug X” are the values for this parameter.

An experiment can be defined by multiple experimental parameters. For instance, the samples could be divided into males and females, and each of these could have ages 1, 2, 5 etc. With this experimental design, there would be 2 experimental parameters, “Gender” and “Age”. “Gender” takes values “male” and “female” and “Age” takes the values “1”, “2” etc.

Experimental parameters and values can be assigned to each sample from the Experiment Grouping link in the workflow browser. These can either be entered manually, or can be imported from a text file, or can be imported from sample attributes. Once these values are provided, you could also shift the parameters from left to right and vice versa. Parameter values within each parameter can also be ordered. All views in **GeneSpring GX** will automatically reflect this order. Suppose you have experimental parameters “Gender” and “Age” and you want your profile plots to show all females first and then all males. Furthermore you would like all females to appear in order of increasing age from left to right and likewise for males. To achieve this, you will need to do the following. First, order the experimental parameters so “Gender” comes first and “Age” comes next. Then order the parameter values for parameter “Gender,” so “Female” comes first and “Male” comes next. Finally, order the parameter values for parameter “Age” so that these are in increasing numeric order.

2.4.5 Conditions and Interpretations

An interpretation defines a particular way of grouping samples into experimental conditions for both data visualization and analysis. When a new experiment is created, **GeneSpring GX** automatically creates a default interpretation for the experiment called “All Samples”. This interpretation just includes all

the samples that were used in the creation of the experiment. New interpretations can be created using the “Create New Interpretation” link in the workflow browser. Once a new interpretation is created, the interpretation will be added to the Interpretations folder within the Navigator.

First, identify the experimental parameters by which you wish to group samples. **GeneSpring GX** will now show you a list of conditions that would result from such grouping. For example, if you choose two parameters, “Gender” and “Age”, and each sample is associated with parameter values Female or Male, and Young or Old, **GeneSpring GX** will take all unique combinations of parameter values to create the following conditions: Female,Old; Female,Young; Male,Old; and Male,Young. Samples that have the same Gender and Age values will be grouped in the same experimental condition. Samples within the same experimental conditions are referred to as “replicates”.

You can choose to ignore certain conditions in the creation of an interpretation. Thus, if you want to analyze only the conditions Female,Old and Female,Young, you can do that by excluding the conditions Male,Old and Male,Young in the creation of the interpretation.

You can also choose whether or not to average replicates within the experimental conditions. If you choose to average, the mean intensity value for each entity across the replicates will be used for display and for analysis when the interpretation is chosen. If you choose not to average, the intensity value for each entity in each sample will be used for display and for analysis when the interpretation is chosen. Such an interpretation is called as Non-averaged interpretation.

Every open experiment has one active interpretation at any given point in time. The active interpretation of each experiment is shown in **bold** in the navigator for that experiment. By default, when an experiment is opened, the “All Samples” interpretation shows active. You can make a different interpretation active, by simply clicking on it in the Navigator. Invoking a view from the View menu will open the view and automatically customize it to the current active interpretation wherever applicable. Most steps in the Workflow browser also take the active interpretation as default and automatically customize analysis to this interpretation, wherever applicable.

An interpretation can be visualized graphically by double-clicking on it. This will launch a profile plot which shows expression profiles corresponding to the chosen interpretation, i.e., the x-axis shows conditions in the interpretation ordered based on the ordering of parameters and parameter values provided in the Experiment Grouping.

Interpretations and Views

Most views in **GeneSpring GX** change their behavior depending on the current active interpretation of the experiment. The table below lists these changes. Refer Table [2.1](#).

View	Behavior on active Interpretation
Scatter Plot Matrix Plot Histogram	Axes show only conditions in this interpretation for averaged interpretations, and individual samples for each condition in the interpretation, for non-averaged interpretations.
Profile Plot Box Whisker Plot	Axes show only conditions in this interpretation for averaged interpretations, and individual samples for each condition in the interpretation, for non-averaged interpretations. Parameter markings are shown on the x-axis.
Venn Diagram	Interpretation does not apply.
Spreadsheet Heat Map	Columns show only conditions in this interpretation for averaged interpretations, and individual samples for each condition in the interpretation, for non-averaged interpretations.
Entity Trees	When constructing entity trees, only conditions in this interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations. When double-clicking on an entity tree object in the Navigator, the conditions corresponding to the current interpretation show in the tree.
Condition Trees	When constructing condition trees, only conditions in the chosen interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations. When double-clicking on a condition tree object in the Navigator, the current interpretation is ignored and the view launches with the interpretation used when constructing the tree. If the conditions of the original interpretation and their associated samples are no longer valid, a warning message to that effect will be shown.
Entity Classification	When constructing entity classifications, only conditions in a chosen interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations. When double-clicking on an entity classification object in the Navigator, the columns corresponding to the current interpretation show up.

Table 2.1: Interpretations and Views

Interpretations and Workflow Operations

Most of the analysis steps in the workflow browser depend on the current active interpretation of the experiment. These dependencies are tabulated below. The steps not mentioned in the table do not depend on the active interpretation. Refer Table 2.2.

Changes in Experiment Grouping and Impact on Interpretations

Note that Experiment Grouping can change via creation of new parameters or edits/deletions of existing parameters and parameter values. Such changes made to Experiment Grouping will have an impact on already-created interpretations. The following cases arise:

Workflow Step	Action on Interpretation
Filter probesets by Expression	Runs on all samples involved in all the conditions in the chosen interpretation; averaging is ignored except for purposes of showing the profile plot after the operation finishes.
Filter probesets by Flags	Runs on all samples involved in all the conditions in the chosen interpretation; averaging is ignored except for purposes of showing the profile plot after the operation finishes.
Significance Analysis	The statistical test options shown depend on the interpretation selected. For instance, if the selected interpretation has only one parameter and two conditions then a T-Test option is shown, if the selected interpretation has only one parameter and many conditions then an ANOVA option is shown, and if the selected interpretation has more than one parameter then a multi-way ANOVA is run; averaging in the interpretation is ignored.
Fold Change	All conditions involved in the chosen interpretation are shown and the user can choose which pairs to find fold change between; averaging in the interpretation is ignored.
GSEA	All conditions involved in the chosen interpretation are shown and the user can choose which pairs to perform GSEA on; averaging in the interpretation is ignored.
Clustering	Only conditions in this interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations.
Find Similar Entities	Only conditions in this interpretation are considered for averaged interpretations, and individual samples for each condition in this interpretation are considered for non-averaged interpretations.
Filter on Parameters	All samples involved in conditions in the chosen interpretation are considered irrespective of whether or not the interpretation is an averaged one. Next, the parameter to be matched is restricted to values on only these samples. Once the calculations have been performed, entities passing the threshold are displayed in a profile plot that reflects the chosen interpretation.
Build Prediction Model	All conditions involved in the chosen interpretation are used as class labels for building a model; averaging in the interpretation is ignored.

Table 2.2: Interpretations and Workflow Operations

- Deleting a parameter: If all parameters used in an interpretation have been subsequently deleted, or even renamed, the interpretation’s behavior defaults to that of the “All Samples” interpretation. If however, only a part of the parameters used in an interpretation have been changed, for e.g., if an interpretation uses parameters Gender and Age, and say, Age has been deleted, then the interpretation behaves as if it was built using only the Gender parameter. If the interpretation had any excluded conditions, they are now ignored. If at a later stage, the Age parameter is restored, the interpretation will again start functioning the way it did when it was first created.
- Change in parameter order: The order of parameters relative to each other can be changed from the Experiment Grouping workflow step. If for e.g., Age is ordered before Gender, then the conditions of an interpretation which includes both Gender and Age, will automatically become Old,Female; Young,Female; Old,Male and Young,Male.
- Deleting a parameter value: The interpretation only maintains the conditions that it needs to exclude.

So, if for example, the parameter value Young is changed to Adolescent, an interpretation on the parameter Age without any excluded conditions will have Adolescent and Old as its conditions. Another interpretation on the parameter Age, that excluded the condition Young will also have as its new conditions - Adolescent and Old.

- Change in order of parameter values: If the order of parameter values is changed, the conditions of the interpretation are also accordingly re-ordered. Thus for parameter Age, if value Young is ordered before Old, the conditions of an interpretation with both Gender and Age, will likewise become Female,Young; Female,Old; Male,Young and Male,Old.

The key point to note is that an interpretation internally only maintains the names of the parameters that it was created with and the conditions that were excluded from it. Based on any changes in the Experiment Grouping, it logically recalculates the set of conditions it represents.

2.4.6 Entity List

An Entity List comprises a subset of entities (i.e., genes, exons, genomic regions, etc.) associated with a particular technology. When a new experiment is created, **GeneSpring GX** automatically creates a default entity list called the “All Entities” entity list. This entity list includes all the entities that the experiment was created with. In most cases, all entities present in the samples loaded into the experiment will also be the same as the entities of the technology associated with the samples. In the case of an Exon Expression experiment however, it contains the Core/Full/Extended transcript cluster IDs depending on which option was chosen to create the experiment. Entity list cannot appear in a pathway experiment.

New entity lists are typically created in **GeneSpring GX** as a result of analysis steps like “Filter probesets by Flags” for example. One could also manually create a new entity list by selecting a set of entities in any of the views and then using the *Create Entity List* toolbar button. Note that entities selected in one view will also show selected in all other views as well.

Existing entity lists can be added to a non-pathway experiment of the same technology via *Search*→*Entity Lists*. Please note that entity lists cannot be cut and pasted across experiments.

Entity lists can be translated implicitly across experiments with possibly differing technologies. Implicit translation happens when you click on an entity list in the analysis navigator of an experiment which is not currently active. Data views in the currently active experiment are restricted to the entities in the above list, after translation is performed silently behind the scenes, possibly via a homologue cross organism map using Entrez gene ID to go across.

This implicit translation works across most experiment types with two notable exceptions.

1. Translation into pathway experiments is not performed

2. Translation of miRNA lists into non-miRNA experiments is not allowed. Translation of gene lists from non-miRNA experiments to miRNA experiments happens implicitly but will lead to nothing being visible. TargetScan translated genelist in miRNA experiments do participate in implicit translation though.

Entity lists alone can be translated explicitly across experiments with possibly differing technologies; this happens on the entity list right click menu and results in a new list. The same exceptions apply as for implicit translation. For more details on how this is executed, refer to the section on [Translation](#).

Every open project has utmost one active entity list at any given point in time. When an experiment of the project is opened, the “All Entities” entity list of that experiment becomes the active entity list of the project. You can make a different entity list active, simply by clicking on it in the Navigator. The user experience key to **GeneSpring GX** is the fact that clicking on an entity list restricts all open views to just the entities in that list, making for fast exploration.

Any entity list in **GeneSpring GX** can be made universally available by right clicking on that list in the navigator and clicking the option 'Mark as Global list'. This list would then appear under 'Global lists' and will be available across experiments in a project and across projects too. Actions possible in 'Global lists' are:

1. Highlight List - Makes it active entity list and does the translation into the active experiment.
2. Inspect List - Brings up Entity List Inspector.
3. Export List - Brings up a window to enable exporting the entity list.
4. Translate List - Within the same project or across projects. Obeys general rules of translation.
5. Share List - Active only for Workgroup version of **GeneSpring GX** .
6. Unmark as Global list - Will remove the list from the 'Global lists'.

2.4.7 Entity Tree, Condition Tree, Combined Tree and Classification

Clustering methods are used to identify co-regulated genes. Trees and classifications are the result of clustering algorithms. All clustering algorithms require a choice of an entity list and an interpretation, and allow for clustering on entities, conditions or both.

Performing hierarchical clustering on entities results in an entity tree, on conditions results in a condition tree and on both entities and conditions results in a combined tree. Performing KMeans or SOM on entities results in a classification, on conditions results in a condition tree, and on both entities and conditions result in a classification and condition tree.

A classification is just a collection of disjoint entity lists. Double-clicking on a classification from the navigator results in the current active view to be split up based on the entity lists of the classification. If the active view does not support splitting up, for e.g., if it is already split, or if it is a Venn Diagram view, etc., then the classification is displayed using split up profile plot views. The classification is displayed according to the conditions in the active interpretation of the experiment. A classification can also be expanded into its constituent entity lists, by right-clicking on the classification and using the *Expand as Entity list* menu item.

Double-clicking on the trees will launch the dendrogram view for the corresponding tree. For entity trees, the view will show all the entities and the corresponding tree, while the columns shown will correspond to the conditions in the active interpretation. For condition trees and combined trees, the same tree as was created will be reproduced in the view. However, it may be that the conditions associated with the samples of the tree are now different, due to changes in the experiment grouping. In this case a warning message will be shown. If any of the samples that were used to create the tree are no longer present in the experiment, after performing a Add/Remove Samples operation for e.g., then an error message will be shown and the tree cannot be launched.

Refer to chapter [Clustering](#) for details on clustering algorithms.

2.4.8 Class Prediction Model

Class prediction methods are typically used to build prognostics for disease identification. For instance, given a collection of normal samples and tumor samples with associated expression data, **GeneSpring GX** can identify expression signatures and use these to predict whether a new unknown sample is of the tumor or normal type. Extending this concept to classifying different types of possibly similar tumors, class prediction provides a powerful tool for early identification and tailored treatment.

Running class prediction involves three steps, validation, training and prediction. The process of learning expression signatures from data automatically is called training. Clearly, training requires a dataset in which class labels of the various samples are known. Performing statistical validation on these signatures to cull out signal from noise is called validation. Once validated these signatures can be used for prediction on new samples.

GeneSpring GX supports four different class prediction algorithms namely, Decision Tree, Neural Network, Support Vector Machine and Naive Bayes. These can be accessed from the “Build Prediction Model” workflow step. Each of these algorithms create a class prediction model at the end of the training. These models can be used for prediction on a potentially different experiment using the “Run Prediction” workflow step.

Refer to chapter [Class Prediction: Learning and Predicting Outcomes](#) for details on the class prediction algorithms.

2.4.9 Script

Python and R scripts can be created and saved in **GeneSpring GX** for performing custom tasks and to easily add and enhance features.

To create a new python script, launch the *Tools* → *Script Editor*, refer the chapter [Writing Scripts in GeneSpring GX](#) on scripting to implement the script, and then save the script using the Save button on the toolbar of the Script Editor. This script can later be invoked on a potentially different experiment by launching a new Script Editor and clicking on the Open toolbar button to search for all existing scripts and load the already saved script.

R scripts can be created and saved similarly using the *Tools* → *R Editor*. Refer to the chapter [Writing Scripts in GeneSpring GX](#) on R scripts for details on the R API provided by **GeneSpring GX**.

2.4.10 Pathway

Pathways can be imported into **GeneSpring GX** from BioPAX files using the “Import BioPAX pathways” workflow step. Pathways in BioPAX Level-2 format is supported. Once imported into the system, pathways can be added to the experiment from the search, or by using the “Find Similar Pathways” functionality.

When a pathway view is opened in an experiment by double-clicking, some of the protein nodes will be highlighted with a blue halo around them. These protein nodes have an Entrez ID that match at least one of the entities of the experiment. The pathway view listens to changes in the active entity list by highlighting the protein nodes that match the entities in that list using Entrez ids. The pathway view is also linked to the selection in other views, and the selected protein nodes show with a green halo by default.


Refer to chapter [Pathway Analysis](#) for details on pathway analysis in **GeneSpring GX**.

2.4.11 Inspectors

All the objects mentioned above have associated properties. Some properties are generic like the name, date of creation and some creation notes, while others are specific to the object, e.g., entities in an entity list. The inspectors of the various objects can be used to view the important properties of the object or to change the set of editable properties associated with the object like Name, Notes, etc.

- The Project Inspector is accessible from *Project* → *Inspect Project* and shows a snapshot of the experiments contained in the project along with their notes.
- The Experiment Inspector is accessible by right-clicking on the experiment and shows a snapshot of

the samples contained in the experiment and the associated experiment grouping. It also has the notes that detail the pre-processing steps performed as part of the experiment creation.

- The Sample Inspector is accessible by double-clicking on the sample in the navigator or by right-clicking on the sample. It shows the experiment the sample belongs to, the sample attributes, attachments and parameters and parameter values from all experiments that it is part of. The name and parameters information associated with the sample are uneditable. Sample attributes can be added/changed/deleted from the inspector, as also the attachments to the sample.
- The Technology Inspector is accessible by right-clicking on the experiment and shows a snapshot of all the entities that belong to the technology. None of the properties of the technology inspector are editable. The set of annotations associated with the entities can be customized using the “Configure Columns” button, and can also be searched for using the search bar at the bottom. Further hyperlinked annotations can be double-clicked to launch a web browser with further details on the entity.
- The Entity List Inspector is accessible by double-clicking on the entity list in the navigator or right-clicking on the entity list. It shows the entities associated with the list, and user attributes if any. It also shows the technology of the entity list and the experiments that it belongs to. The set of displayed annotations associated with the entities can be customized using the “Configure Columns” button, and can also be searched for using the search bar at the bottom. Further, entities in the table can be double clicked to launch the Entity Inspector.
- The Entity Inspector is accessible in the following ways:
 - Double clicking on an entity in the entity list inspector described above
 - Double clicking on some of the views like Scatter plot, MvA plot, Profile plot, Heat map
 - Selecting an entity in any view and clicking on the ‘Inspect entity’  icon toolbar button.
 - By using the key binding Ctrl-I or by using the menu *View* → *Inspect Entities*

The entity inspector window shows the Id and the technology relevant to the selected entity. The inspector also has tabs to view the following: beginenumerate

- *Annotation*: Lists annotations. The set of default annotations associated with the entity can be customized by using the “Configure Columns” button at the bottom.
- *Data*: Shows the raw and normalized data associated with the entity in all the samples of the experiment, along with the flag.
- *Box whisker plot*: With the normalized data under the current active interpretation
- *Profile Plot*: With the normalized data under the current active interpretation. endenumerate
- Inspectors for Entity Trees, Condition Trees, Combined Trees, Classifications, Class Prediction Models are all accessible by double-clicking or right-clicking on the object in the navigator, and provide basic information about it. The name and notes of all these objects can be changed from the inspector.

2.4.12 Hierarchy of objects

All the objects described above have an inherent notion of hierarchy amongst them. The project is right at the top of the hierarchy, and is a parent for one or more experiments. Each experiment is a parent for one or more samples, interpretations and entity lists. Each entity list could be a parent for other entity lists, trees, classifications, class prediction models, pathways, or folders containing some of these objects. The only exceptions to this hierarchy are technologies and scripts that do not have any parentage.

Additionally, many of these objects are *first class objects* that can exist without any parent. This includes experiments, entity lists, samples, class prediction models and pathways. Interpretations, trees and classifications, however cannot exist independently without their parents. Finally, the independent objects can have more than one parent as well. Thus an experiment can belong to more than one project, samples can belong to more than one experiment and so on.

Note that in the case of independent objects, only those that do have a valid parent show up in the navigator. However all objects with or without parents show up in search results.

2.4.13 Right-click operations

Each of the objects that show up in the navigator have several right-click operations. For each object, one of the right-click operations is the default operation and shows in bold. This operation gets executed if you double-click on the object.

The set of common operations available on all objects include the following:

- Inspect object : Most of the objects have an inspector that displays some of the useful properties of the object. The inspector can be launched by right-clicking on the object and choosing the inspect object link.
- Share object : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation can be used to share the object with other users of the **GeneSpring GX** workgroup.
- Change owner : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation can be used by a group administrator to change the owner of the object.

The other operations available on each of the objects are described below:

Experiment

- Open Experiment : (default operation) This operation opens the experiment in **GeneSpring GX**. Opening an experiment opens up the experiment navigator in the navigator section of **GeneSpring GX**. The navigator shows all the objects that belong to the experiment, and the desktop shows the views of the experiment. This operation is enabled only if the experiment is not already open.
- Close Experiment : This operation closes the experiment, and is enabled only if the experiment is already open.
- Inspect Technology : This operation opens up the inspector for the technology of the experiment.
- Create New Experiment : This operation can be used to create a copy of the chosen experiment. The experiment grouping information from the chosen experiment is carried forward to the new experiment. In the process of creating the copy, some of the samples can be removed, or extra samples can be added if desired.
- Remove Experiment : This operation removes the experiment from the project. Note that the remove operation only disassociates the experiment with this project. The experiment could still belong to other projects in the system, or it could even not belong to any project.
- Delete Experiment : This operation will permanently delete the experiment from the system. All the children of the experiment will also be permanently deleted, irrespective of whether they are used in other experiments or not. The only exception to this is samples. So, if an experiment contains ten samples, two of which are used in another experiment, this operation will result in deleting all the eight samples that belong only to this experiment. The remaining two samples will be left intact.

Sample

- Inspect Sample : (default operation) This will open up the inspector for the sample.
- Download Sample : This operation enables downloading the sample to a folder of choice on the local file system.

Samples Folder

- Add Attachments : This operation can be used to upload attachments to all the samples in the folder. Multiple files can be chosen to be added as attachments. **GeneSpring GX** checks the files to see if the name of any of the file (after stripping its extension) matches the name of any sample (after stripping its extension) and uploads that file as an attachment to that sample. Files that do not match this condition are ignored. Note that if a file without a matching name needs to be uploaded as an attachment, it can be done from the sample inspector.
- Add Attributes : This operation can be used to upload sample attributes for all the samples in the folder. **GeneSpring GX** expects a comma or tab separated file in the following tabular format. The first column of the file should be the name of the samples. All the remaining columns will be considered as sample attributes. The column header of each column is taken as the names of the

sample attribute. Each cell in this tabular format is assigned as the value for the corresponding sample (row header) and sample attribute (column header).

- Download Samples : This operation can be used to download all the raw files of the samples in bulk to a folder of choice on the local filesystem.

Interpretation

- Open Interpretation : (default operation) This opens a profile plot view of the interpretation.
- Edit Interpretation : This allows for editing the interpretation. The parameters of the interpretation, conditions to exclude, name and notes can all be edited.
- Delete Interpretation : This operation deletes the interpretation from the experiment. Note that there is no notion of removing an interpretation, since an interpretation is not an independent object and always exists only within the experiment.

Entity List

- Highlight List : This operation restricts all the views in the experiment to the entities of the chosen list.
- Inspect List: This launches the entity list inspector. For more details, refer to [Inspectors](#).
- Export List : This operation can be used to export the entity list and associated data and annotations as a plain text file. One can choose an interpretation according to which the raw and normalized data will be exported, if chosen. If the experiment has flags, then can also choose to export the flags associated with the entities of this list. If the entity list has data associated with it as a result of the analysis using which the list was created, these can also be exported. Finally, one can also choose which annotations to export with the entity list.
- Copy List : This allows the copying of the entity list into the **My Favorites** folder.
- Remove List : This operation removes the entity list from the experiment. Note that the remove operation only disassociates this entity list and all its children with the experiment, and does not actually delete the list or its children. The entity list and its children could still belong to other experiments in the system, or they may even exist independently without belonging to any experiment.
- Delete List : This operation will permanently delete the list and all its children from the system.

Entity List Folder

- Rename Folder : This operation can be used to rename the folder.

- Remove Folder : This operation will remove the folder and all its children from the experiment. Note that the remove operation will delete the folder itself, but will only disassociate all the children from the experiment. The children could still belong to zero or more experiments in the system.
- Delete Folder : This operation will permanently delete the folder and all its children from the system.

Classification

- Open Classification : (default operation) This operation results in the current active view to be split up based on the entity lists of the classification. If the active view does not support splitting up, for e.g., if it is already split, or if it is a Venn Diagram view, etc., then the classification is displayed using split up profile plot views.
- Expand as Entity List : This operation results in creating a folder with entity lists that each correspond to a cluster in the classification.
- Delete Classification : This operation will permanently delete the classification from the experiment. Note that there is no notion of removing a classification, since a classification is not an independent object and always exists only within the experiment.

Entity/Condition/Combined Tree

- Open Tree : (default operation) This operation opens up the tree view for this object. In the case of entity trees, the tree shows columns corresponding to the active interpretation. In the case of condition and combined trees, the tree shows the conditions that were used in the creation of the tree.
- Delete Tree : This operation will permanently delete the tree from the experiment. Note that there is no notion of removing a tree, since a tree is not an independent object and always exists only within the experiment.

Class Prediction Model

- Remove Model : This operation removes the model from the experiment. Note that this operation only disassociates the model with the experiment and does not actually delete the model. The model could still belong to other experiments in the system, or may even exist without being part of any other experiment.
- Delete Model : This operation permanently deletes the model from the system.

Pathway

- Open Pathway : (default operation) This operation opens up the pathway view. Protein nodes in the pathway view that have an Entrez id matching with an entity of the current experiment have a

blue halo around them.

- Remove Pathway : This operation removes the pathway from the experiment. Note that this operation only disassociates the pathway with the experiment and does not actually delete the pathway. The pathway could still belong to other experiments in the system, or may even exist without being part of any other experiment.
- Delete Pathway : This operation permanently deletes the pathway from the system.

2.4.14 Search

An instance of **GeneSpring GX** could have many projects, experiments, entity lists, technologies etc. All of these carry searchable annotations. **GeneSpring GX** supports two types of search - a simple keyword search and a more advanced condition based search. Search in **GeneSpring GX** is case insensitive. The simple keyword search searches over all the annotations associated with the object including its name, notes, etc. Leaving the keyword blank will result in all objects of that type being shown in the results. The advanced condition based search allows performing search based on more complex search criteria joined by OR or AND conditions, for e.g., search all entity lists that contain the phrase “Fold change” and created after a certain date. The maximum number of search results to display is set at 100 and can be changed in the box provided in the **Search Parameters** wizard(step1). It can also be changed from **Tools** → **Options** → **Miscellaneous** → **Search Results**.

Depending on the type of object being searched for, a variety of operations can be performed on results of the search. The **Search Results** wizard (step3) displays a message about the total number of results obtained for that search as well as the number of results on that page. The total number of pages are also given and the user can navigate to the page of his/her choice by entering the page number in the box provided. All the toolbar buttons on the search results page operate on the set of selected objects in the result.

Search Experiments

- Inspect experiments : This operation opens up the inspector for all the selected experiments.
- Delete experiments : This operation permanently deletes the selected experiments and their children from the system. The only exception to this is samples, and samples will be deleted only if they are not used by another experiment in the system. If the experiment being deleted also belongs to the currently open project and it is currently open, it will be closed and will show with a grey font in the project navigator. Also, at a later stage, on opening a project that contains some of these deleted experiments, the experiments will show in grey in the navigator, as a feedback of the delete operation.
- Add experiments to project : This operation adds the selected experiments to the current project, if one is open. If any of the selected experiments already belong to the project, then they are ignored.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation allows sharing the experiment with other users of the workgroup.

Search Samples

- Inspect samples : This operation opens up the inspector for all the selected samples.
- Create new experiment : This operation creates a new experiment with the set of selected samples. If the selected samples do not belong to the same technology an error message will be shown. This operation will close the search wizard and launch the new experiment creation wizard with the set of selected samples.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation allows sharing the samples with other users of the workgroup.
- View containing experiments : This operation shows a dialog with the list of experiments that the selected samples belong to. This dialog also shows an inverse view with the list of all samples grouped by the experiments that they belong to. One can select and add experiments to the current project from this view.

Search Entity Lists

- Inspect entity lists : This operation opens up the inspector for all the selected entity lists.
- Delete entity lists : This operation will permanently delete the selected entity lists from the system. Note that only the selected entity lists will be deleted, and if they belong to any experiments, their children in each of those experiments will remain intact. If the entity lists being deleted belong to one or more of the currently open experiment, the navigator of the experiment will refresh itself and the deleted entity lists will show in grey.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation allows sharing the entity lists with other users of the workgroup.
- View containing experiments : This operation shows a dialog with the list of experiments that the selected entity lists belong to. This dialog also shows an inverse view with the list of all entity lists grouped by the experiments that they belong to. One can select and add experiments to the current project from this view.
- Add entity lists to experiment : This operation adds the selected entity lists to the active experiment. The entity lists get added to a folder called “Imported Lists” under the All Entities entity list. Entity lists that do not belong to the same technology as the active experiment are ignored.

Search Entities

The search entities wizard enables searching entities from the technology of the active experiment. The first page of the wizard allows choosing the annotations to search on, and the search keyword. The second page of the wizard shows the list of entities that match the search criterion. A subset of entities can be selected here to create a custom list. On clicking next and then finish, an entity list gets created with all the entities that match the search criterion. This entity list is added under the All Entities entity list.

Search Pathways

- Inspect pathways : This operation opens up the inspector for all the selected pathways.
- Delete pathways : This operation will permanently delete the selected pathways from the system. If the pathways being deleted belong to one or more of the currently open experiment, the navigator of the experiment will refresh itself and the deleted pathways will show in grey. Also, at a later stage, on opening an experiment that contains some of these deleted pathways, the pathways will show in grey in the navigator, as a feedback of the delete operation.
- Add pathways to experiment : This operation adds the selected pathways to the active experiment. The pathways get added to a folder called “Imported Pathways” under the All Entities entity list.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation allows sharing the pathways with other users of the workgroup.

Search Prediction Models

- Inspect models : This operation opens up the inspector for all the selected models.
- Delete models : This operation will permanently delete the selected models from the system. If the models being deleted belong to one or more of the currently open experiment, the navigator of the experiment will refresh itself and the deleted models will show in grey. Also, at a later stage, on opening an experiment that contains some of these deleted models, the models will show in grey in the navigator, as a feedback of the delete operation.
- Add models to experiment : This operation adds the selected models to the active experiment. The models get added to a folder called “Imported Models” under the All Entities entity list. Models that do not belong to the same technology as the active experiment are ignored.

Search Scripts

- Inspect scripts : This operation opens up the inspector for all the selected scripts.
- Delete scripts : This operation will permanently delete the selected scripts from the system.
- Open scripts : This operation opens the selected scripts in Python or R Script Editor in the active experiment.

Search Technology

- Inspect technologies : This operation opens up the inspector for all the selected technologies.

Search All

GeneSpring GX provides the ability to search for multiple objects at the same time using the Search All functionality.

- Inspect objects : This operation opens up the inspector for all the selected objects.
- Delete objects : This operation will permanently delete the selected objects from the system. Samples that belong to any experiment will not be deleted.
- Change permissions : This operation is disabled in the desktop mode of **GeneSpring GX**. In the workgroup mode, this operation allows sharing the objects with other users of the workgroup.

2.4.15 Saving and Sharing Projects

The state of an open project, i.e., all experiments and their respective navigators, are always auto-saved and therefore do not need to be saved explicitly. This is however not true of the open views, which unless saved explicitly are lost on shutdown. Explicit saving is provided via a *Save Current View* link on the Workflow browser.

What if you wish to share your projects with others or move your projects from one machine to another?

Projects can be shared with other users using the **Export Project Zip** functionality from *Project* → *Export/Import project zip*. This zip file is portable across platforms i.e., Linux, Windows, Mac etc.

Export Project Zip - This feature allows the user to export project as a whole along with experiments, in a zip format. Some/all experiments within a project can be exported. When a project zip is created, in case of Generic Single Color, Generic Two Color and the experiments migrated from GX7, the technologies are bundled along with the zip file. The standard technologies and Affymetrix Custom technologies are not bundled along with the zipped project. These can be selected from the **Choose Technologies** window that appears after the experiments to be exported are chosen. The zipped projects are imported by the second user using the **Import Project Zip** feature. This allows the import of the zipped project along with the experiments. In case standard technologies were not packaged with the project zip, a message will be prompted asking to download the technologies needed to open the project and experiments. For Affymetrix Custom experiments, the technology will have to be created prior to importing zipped projects in case the technology associated with the custom experiment is not exported with the project zip.

2.4.16 Software Organization

At this point, it may be useful to provide a software architectural overview of **GeneSpring GX**. **GeneSpring GX** contains three parts, a UI layer, a database and a file system. The file system is where all

objects are stored physically; these are stored in the app/data subfolder in the installation folder. A Derby database carries all annotations associated with the various objects in the file system (i.e., properties like notes, names etc which can be searched on); a database is used to drive fast search. Finally, the UI layer displays relevant objects organized into projects, experiments, analysis etc.

2.5 Exporting and Printing Images and Reports

Each view can be printed as an image or as an HTML file: Right-Click on the view, use the Export As option, and choose either Image or HTML. Image format options include jpeg (compressed) and png (high resolution).

Exporting Whole Images. Exporting an image will export only the VISIBLE part of the image. Only the dendrogram view supports whole image export via the Print or Export as HTML options; you will be prompted for this. The Print option generates an HTML file with embedded images and pops up the default HTML browser to display the file. You need to explicitly print from the browser to get a hard copy.

Finally, images can be copied directly to the clipboard and then pasted into any application like PowerPoint or Word. Right-Click on the view, use the *Copy View* option and then paste into the target application. Further, columns in a dataset can be exported to the Windows clipboard. Select the columns in the spreadsheet and using Right-Click *Select Columns* and then paste them into other applications like Excel using Ctrl-V.

2.6 Scripting

GeneSpring GX has a powerful scripting interface which allows automation of tasks within **GeneSpring GX** via flexible Jython scripts. Most operations available on the **GeneSpring GX** UI can be called from within a script. To run a script, go to *Tools* → *Script Editor*. A few sample scripts are packaged with the demo project. For further details, refer to the [Scripting](#) chapter. In addition, R scripts can also be called via the *Tools* → *R Script Editor*.

2.7 Options

Various parameters about **GeneSpring GX** are configurable from *Tools* → *Options*. These include algorithm parameters and various URLs.

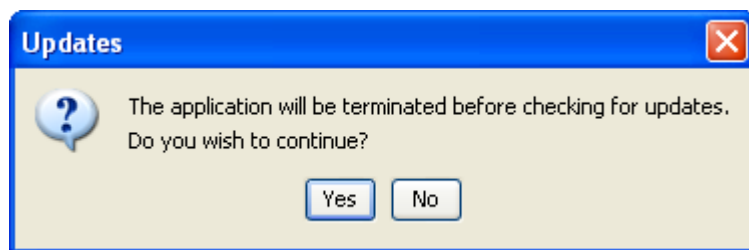


Figure 2.5: Confirmation Dialog

2.8 Update Utility

GeneSpring GX has an update utility that can be used to update the product or get data libraries needed for creating an experiment. These data library updates and product updates are periodically deployed on the **GeneSpring GX** product site and is available online through the tool. The update utility is available from the *Annotations* and *Help* → *Update Product*. This will launch the update utility that will contact the online update server, verify the license, query the sever and retrieve the update (if any) that are available. Note that you have to be connected to the Internet and should be able to access the **GeneSpring GX** update server to fetch the updates. In situations where you are unable to connect to the update server, you can do an update form a file provided by Agilent support.

2.8.1 Product Updates

GeneSpring GX product updates are periodically deployed on the update server. These updates could contain bug fixes, feature enhancements and product enhancements. Choosing product update from *Help* → *Update Product* → *from Web* will prompt a dialog stating that the application will be terminated before checking for updates. Confirm to close the application. This will launch the update utility that will contact the online update server, verify the license, query the sever and retrieve the product update (if any) available. See Figure 2.5

If updates are available, the dialog will show the available updates. Left-Click on the check box to select the update. If multiple updates are available, you can select multiple updates simultaneously. Details about the selected update(s) will be shown in the description box of the update dialog. Left-Click *OK* will download the update and execute the update to apply it on your product. When you launch the tool, these updates will be available. To verify the update, you can check the version of build number from the *Help* → *About GeneSpring GX* . See Figure 2.6

2.9 Getting Help

Help is accessible from various places in **GeneSpring GX** and always opens up in an HTML browser.

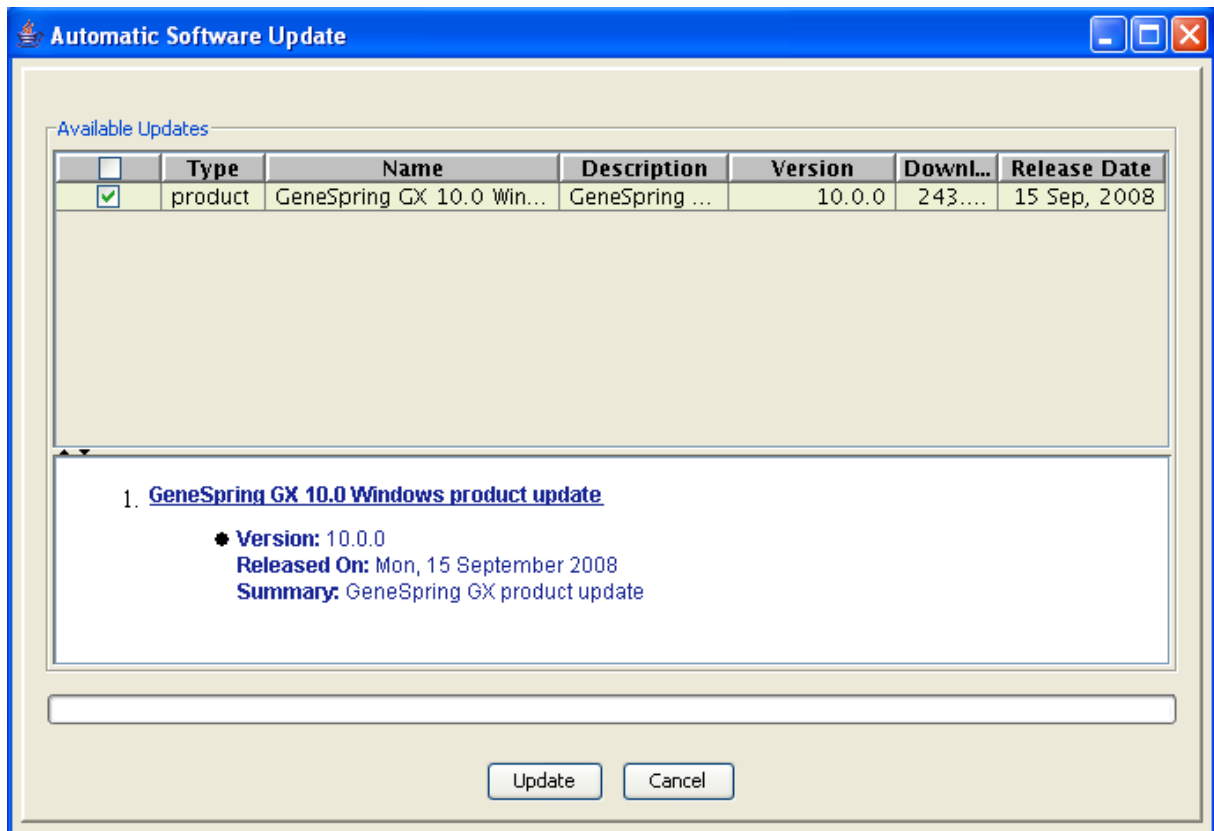


Figure 2.6: Product Update Dialog

Single Button Help. Context sensitive help is accessible by pressing F1 from anywhere in the tool.

All configuration utility and dialogs have a *Help* button. Left-Click on these takes you to the appropriate section of the help. All error messages with suggestions of resolution have a help button that opens the appropriate section of the online help. Additionally, hovering the cursor on an icon in any of the windows of **GeneSpring GX** displays the function represented by that icon as a tool tip.

Help is accessible from the drop down menu on the menubar. The Help menu provides access to all the documentation available in **GeneSpring GX**. These are listed below:

- **Help:** This opens the Table of Contents of the on-line **GeneSpring GX** user manual in a browser.
- **Documentation Index:** This provides an index of all documentation available in the tool.
- **About GeneSpring GX :** This provides information on the current installation, giving the edition, version and build number.

Chapter 3

Technology and Biological Genome

3.1 Technology

Technology in **GeneSpring GX** is defined as the package of data regarding array design, biological and other information about the entities, eg., Entrez gene ID, GO accession etc. Technology is available for each individual array type-i.e., the technology for Affymetrix HG-U133_plus.2 would contain information specific to its design and would thus differ from other technologies, like the Agilent 12097 (Human 1A). An experiment comprises samples which all belong to the same technology.

A technology initially must be installed for each new array type to be analyzed. For standard arrays from Affymetrix, Agilent and Illumina, technologies have been created beforehand and **GeneSpring GX** will automatically prompt for downloading these technologies from Agilent's server whenever required. For other array types, technologies can be created in **GeneSpring GX** via the custom technology creation wizard from *Annotations*→*Create Technology*. See Figure 3.1

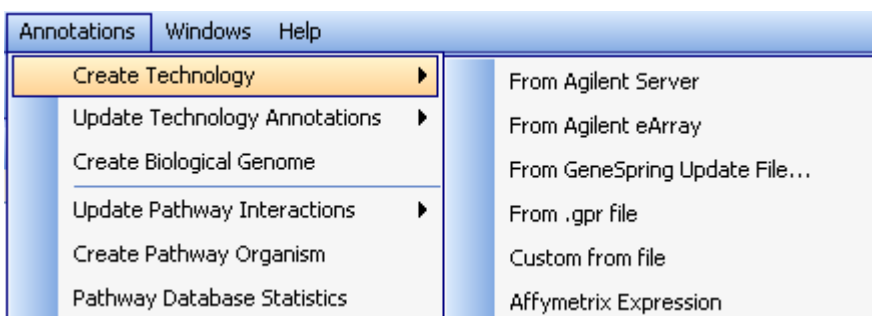


Figure 3.1: Create Technology

3.1.1 Standard Technology Creation

The creation of a Standard Technology involves processing of the information present in the annotation files into a standard internal format used in **GeneSpring GX** . This is done for greater efficiency while using functionalities such as GO Analysis.

The different files used for different technologies are detailed below:

- **Affymetrix Expression:**

The files that are used for creating a Standard Technology are .csv, .cdf, .psi, .cif and probetab. The .cif file is needed when summarization is being performed using MAS5. Likewise the probetab file is required while doing summarization using GCRMA. Additional parsing of the data files goes on during technology creation, for eg., the chromosomal information pertaining to a probe (number, strand, position and orientation) in the Affymetrix annotation file is present in a single column and during the process of technology creation; this is split into 4 different columns. The information required for creating a Standard Technology is taken from the following site: <http://www.affymetrix.com/analysis/index.affx>

- **Affymetrix Exon Expression:**

The files that are used for creating a Standard Technology are .clf, .pgf and the transcript level .csv annotation file. The meta probeset lists regarding the Core, Extended and Full transcripts are the same as Affymetrix files and are packaged with the Standard Technology. The information required for creating a Standard Technology is taken from the following website: <http://www.affymetrix.com/analysis/index.affx>.

- **Affymetrix Exon Splicing:**

The files that are used for creating a Standard Technology are .clf, .pgf, probeset level .csv annotation file and the transcript level .csv annotation file. The meta probeset and the probeset files regarding the Core, Extended and Full transcripts and exons are the same as Affymetrix files and are packaged with the Standard Technology. The information required for creating a Standard Technology is taken from the following website: <http://www.affymetrix.com/analysis/index.affx>.

- **Illumina:**

The creation of a Standard Technology for Illumina arrays uses the information content of the .bgx manifest file to associate the annotations with the probes. The information required for creating a Standard Technology is taken from the following website: <http://www.switchtoi.com/annotationfiles.ilmn>.

- **Agilent Single and Two Colour:**

The creation of a Standard Technology for Agilent arrays involves parsing the biological information present in the annotation file into a **GeneSpring GX** recognizable format. For eg., the chromosomal information pertaining to a probe (chromosome number, strand, position and orientation) is present in a single column in the annotation file while the GO annotations are present in 3 columns. During technology creation, the chromosomal information is parsed into 4 columns while the GO annotations are collapsed into 1 column. Annotations for Agilent arrays are available on the following website: <http://www.chem.agilent.com>

- **Agilent miRNA:**

The technology creation is done spontaneously for this experiment type and is referred to as *technology creation on the fly*. For more details refer to section on [Technology creation on the fly](#). As and when annotation files become available, Standard Technologies will be created and can be downloaded from the update server.

- **Real Time PCR:**

The technology creation is dependant on the samples given and each individual experiment has its own technology. This technology creation does not have annotations associated with it. The user can update annotations after experiment creation from **Utilities**→*Update RTPCR Technology Annotations* under the workflow navigator.

- **Copy Number Analysis:**

For Copy Number Analysis, **GeneSpring GX** 11.0 supports the following standard technologies:

1. **Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set**
2. **Affymetrix Human Mapping 100K Set**
3. **Illumina Genotyping output files from GenomeStudio**

Refer to Chapter [Copy Number Analysis](#) for details.

- **Association Analysis:**

GeneSpring GX supports the following technologies for Association Analysis experiments:

- Affymetrix Mapping 50K Xba240
- Affymetrix Mapping 50K Hind240
- Affymetrix Mapping 50K Xba240 and 50K Hind240
- Affymetrix Mapping 250K Nsp
- Affymetrix Mapping 250K Sty
- Affymetrix Mapping 250K Nsp and 250K Sty
- Affymetrix GenomeWide SNP5
- Affymetrix GenomeWide SNP6
- Genotyping Output files from Illumina GenomeStudio
- Any file created in Illumina GenomeStudio output format (refer to [Illumina File Format](#) section for details).

Refer to [Technology](#) section for details.

3.1.2 Agilent eArray Technology Creation

Agilent Single color and Two color technologies can also be created for arrays ordered through the eArray portal of Agilent. This can be accessed from *Annotations*→*Create Technology*→*From Agilent*

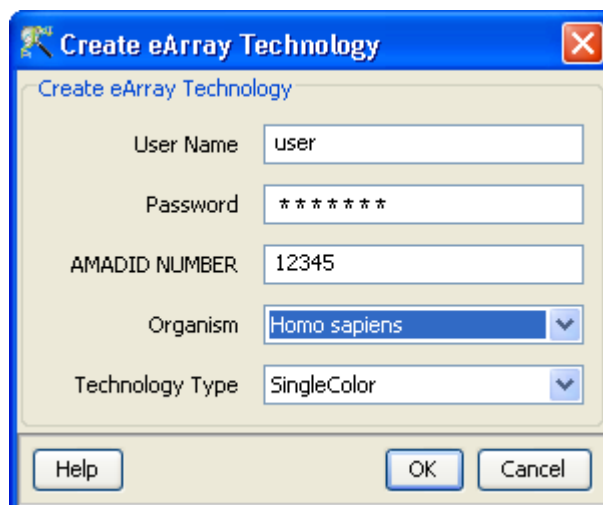


Figure 3.2: Technology Creation

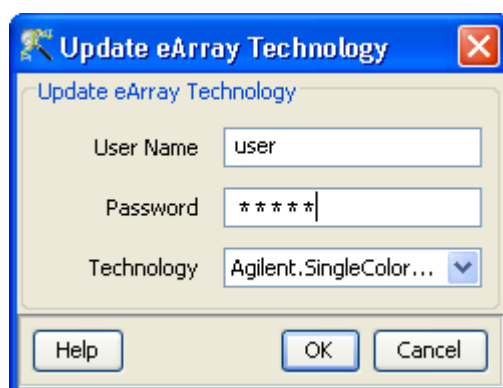


Figure 3.3: Technology Update

eArray. An account at eArray is required to create technology using this feature. Details such as user name, password for your eArray account, AMADID Number, organism and technology type are required for creating an eArray technology. See Figure 3.2. Once the details are provided, a technology is created along with annotation columns like Entrez-ID, GO etc (if available) from the tab delimited text (TDT) file of the specified eArray.

If the user wants to populate the created technology with more annotations, then this can be done through *Annotations*→*Update Technology Annotations*→*From Agilent eArray*. This opens up a window in which the user needs to key in information regarding the User Name, Password and the Technology Name. See Figure 3.3. This opens up a window which is similar to the step 3 of [Update Technology Annotations](#) from file. **GeneSpring GX** uses the information present in the 'AllAnnotations' file of the eArray to populate the technology with further annotations. In case this file is not available for the eArray, then it uses the TDT file to populate the technology.

3.1.3 Custom Technology Creation

GeneSpring GX allows the user to create a custom technology. This is useful in cases where the user has a custom array from the vendors mentioned above or has an array from a different vendor. The option to create a custom technology Generic One or Two Color arrays can be availed from *Annotations*→*Create Technology*→*Custom from file*. Custom Technology for Affymetrix Expression arrays (if a custom .cdf file is available) can be created using *Annotations*→*Create Technology*→*Affymetrix Expression*. For GenePix Results (.gpr) format files use *Annotations*→*Create Technology*→*From .gpr file*.

You can create a Custom Technology to run an Illumina Association Analysis experiment on any file created in GenomeStudio output format (refer to [Illumina File Format](#) section for details).

3.1.4 Technology creation on the fly

This option is used by the application when Agilent FE files are used to create an experiment and the technology for the FE file does not exist either in the **GeneSpring GX** application or on the Agilent server. It let's the user proceed with experiment creation and a technology is created with just the identifier column without any annotations. The annotations can be updated later on as and when the annotation files are available. This can be done from *Annotations*→*Update Technology Annotations*. This update can be done using either the *From Agilent eArray*(Refer to section on [eArray](#)) or the *From file or Biological Genome* options.

An organism is needed for creating the technology and the user is prompted for the same during the workflow. Please note that technology creation on the fly will also come into picture when the technology does not exist in **GeneSpring GX** and the application could not connect to the Agilent server to download for the technology.

3.1.5 Inspection of Technology

A technology once created or downloaded can be inspected at any time using the Technology Inspector. It is accessible by right-clicking on the experiment name in the project navigator and provides information regarding the organism, type (Single or Two Color), version (for Standard Technology) and the number of entities and the date of creation. Except for the organism name and notes, none of the other information can be edited. The set of annotations associated with the entities can be customized using the “Configure Columns” button, and can also be searched for using the search bar at the bottom. Further hyperlinked annotations can be double-clicked to launch a web browser with further details on the entity.

3.1.6 Technology Deletion

Technologies once created can be deleted if no longer in use. This can be done using *Search* → *Technology*. The toolbar in the search wizard has an icon for deleting technology.

3.2 Update Technology

The available technologies in **GeneSpring GX** can be updated regularly. Updates can be carried out in a file-based format by using the necessary file (provided by **GeneSpring GX** support on request) or can be updated from the update server. Updates are available on the server whenever new data libraries are made available by the chip manufacturers.

Data libraries are also required for other applications in the tool. For example, the Genome Browser would require different kinds of track data for different organisms to display the analysis results on the organism's genome. Gene Ontology (GO) data is necessary for GO analysis. To see the available updates, go to *Annotations* → *Update Technology Annotations* → *From Agilent Server*. This will contact the update server, validate the license and show the data libraries available for update. Select the required libraries by Left-Click on the check box next to the data library. Details of the selected libraries will appear in the text box below the data library list. See Figure 3.4

You can Left-Click on the check box header to select or unselect all the data libraries. Left-Click on a check box will toggle the selection. Thus if the check box is unselected, Left-Click on it will select the row. If the row is selected, Left-Click on the check box will unselect the row. Shift-Left-Click on the check box will toggle the selection of all rows between the last Left-Click and Shift-Left-Click .

You can sort the data library list on any column by Left-Click on the appropriate column header.

3.2.1 Automatic Query of Update Server

When experiments are created, if the appropriate libraries are not available, the tool will prompt the user to download the required data library before proceeding further. See Figure 3.5

3.2.2 Update Technology Annotations

Update Technology Annotations, enables the user to update the annotations of an existing Standard or Generic technology. It is a particularly useful feature when newer information necessitates updating an existing technology. The Standard Technologies can typically be updated from the web whenever the chip vendor releases newer annotation. However, this particular feature of **Update Technology Annotations**

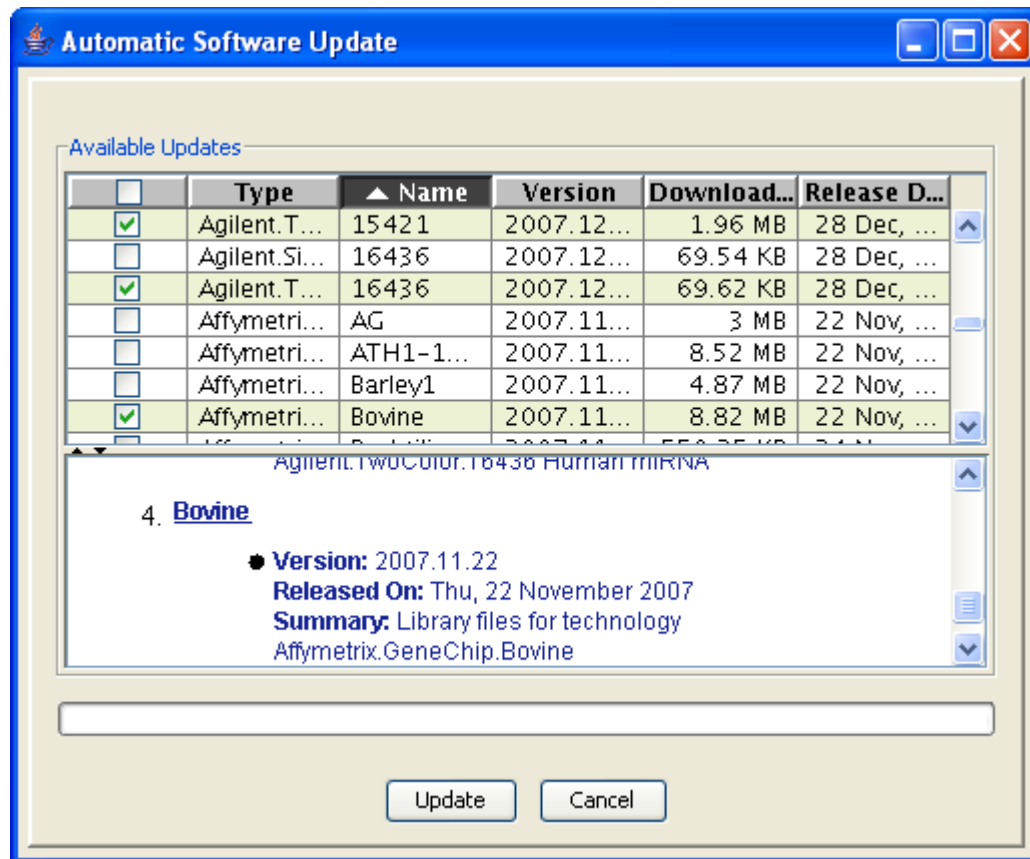


Figure 3.4: Data Library Updates Dialog

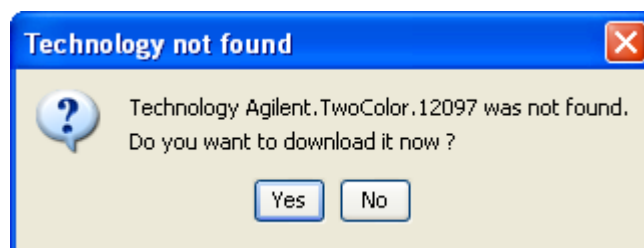


Figure 3.5: Automatic Download Confirmation Dialog

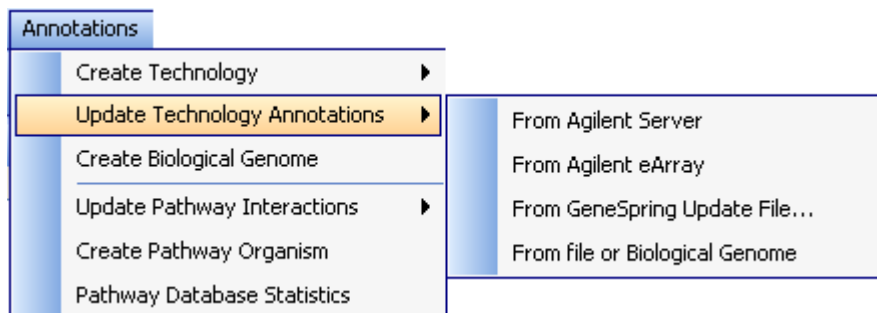


Figure 3.6: Update Technology Annotations

for Standard Technologies is used when you would want to add additional information over and above that provided by the vendor. Different ways to update technology annotations can be accessed from the menu *Annotations* → *Update Technology Annotations*. See Figure 3.6

1. **From Agilent Server** - Updates are available on the server whenever new data libraries are made available by the chip manufacturers.
2. **From Agilent eArray** - Agilent technologies can be updated from eArray directly. You will require username and password of eArray to access.
3. **From GeneSpring Update File** - Standard technologies can be updated using GeneSpring Update file (provided by GeneSpring GX support on request)
4. **From file or Biological Genome** - It can be accessed from the menu *Annotations* → *Update Technology Annotations* → *From file or Biological Genome*:
 - (a) **Step 1 of 3** - Here the user specifies the technology as well as the source from which it has to be updated. The technology can be updated either from a file or from the **Biological Genome** of that organism. If the **Biological Genome** of that organism does not exist, then the user can create a genome from *Annotations* → *Create Biological Genome*. For more details on the creation of a genome, refer to [Biological Genome](#). If the user chooses to update from a file, then it should be chosen accordingly via the *Choose file* option. The file from which the update is to be performed has to be in a tabular format. This is seen in Figure 3.7.
 - (b) **Step 2 of 3** - This step appears only if the update source is a file. This step asks the user to input the file format of the annotations update file. This involves specifying format options, i.e., the Separator, Text qualifier, Missing value indicator and Comment Indicator of the file. This is seen in Figure 3.8.
 - (c) **Step 3 of 3** - The annotation columns are merged with the existing technology using a technology identifier. This step asks the user to specify the identifier and to choose the column to be updated from the annotation file/genome. While specifying the columns, column marks should be assigned (similar to how it was done while creating the Generic technology). It is recommended that the user chooses a column with unique values (Ex:Entrez-ID) as the identifier. Three kinds of updates are possible:
 - Append to the existing information,

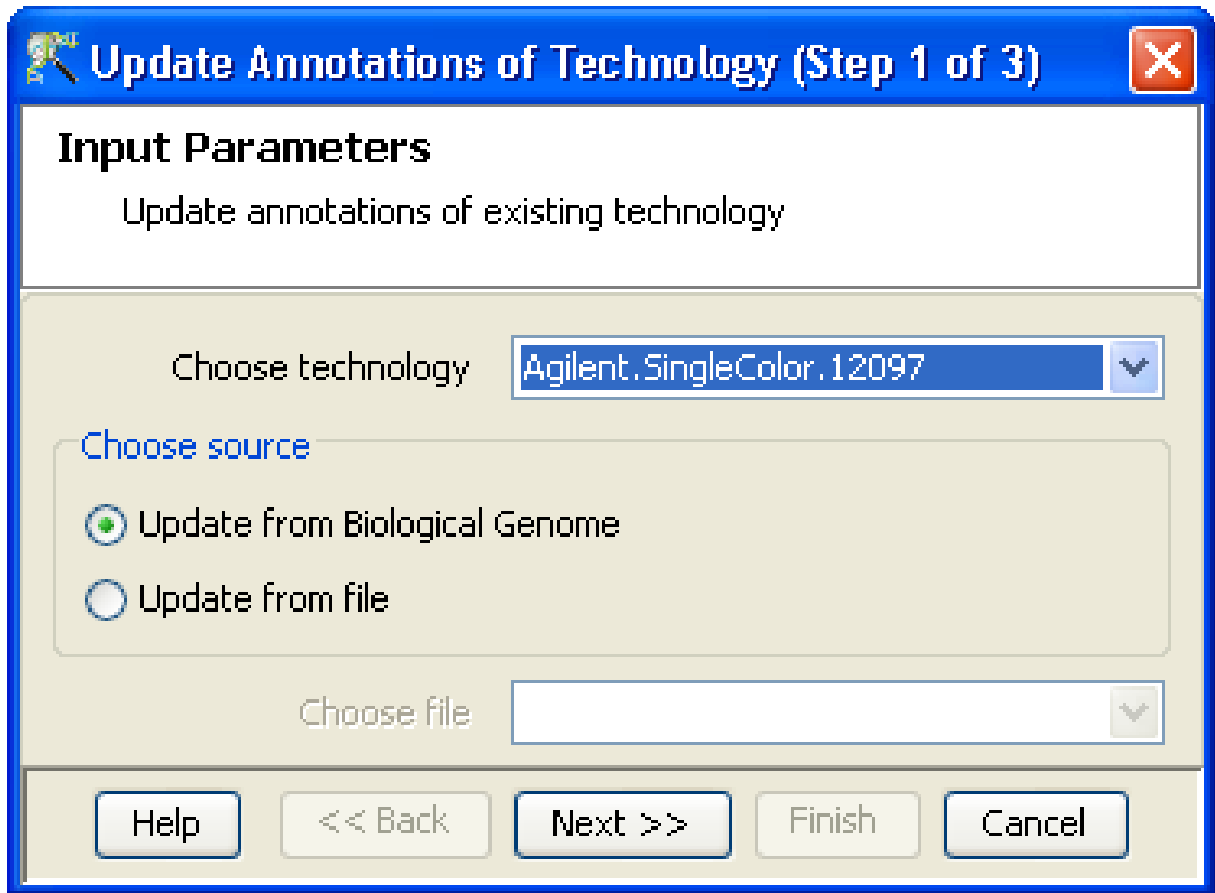


Figure 3.7: Input Parameters

- Overwrite
- Fill in the missing values.

Appending the values will retain the original value as well as add the new value. Overwrite will replace the original value with the newer one, whereas fill in missing values will add values at places where previously there were none. This is seen in Figure 3.9.

The updated annotation values for existing columns can be seen by right click on *Experiment*→*Inspect Technology*.

3.3 Translation

Translation is a feature that allows comparison of entity lists between experiments of different technologies. A standard use case of translation involves comparison of experiments done on a single organism but different technologies, e.g., Human samples on HG_U95Av2 and HG-U133_Plus_2. Another situation would be to identify the homologues, eg mapping Human genes to Mouse genome. The automated detection of

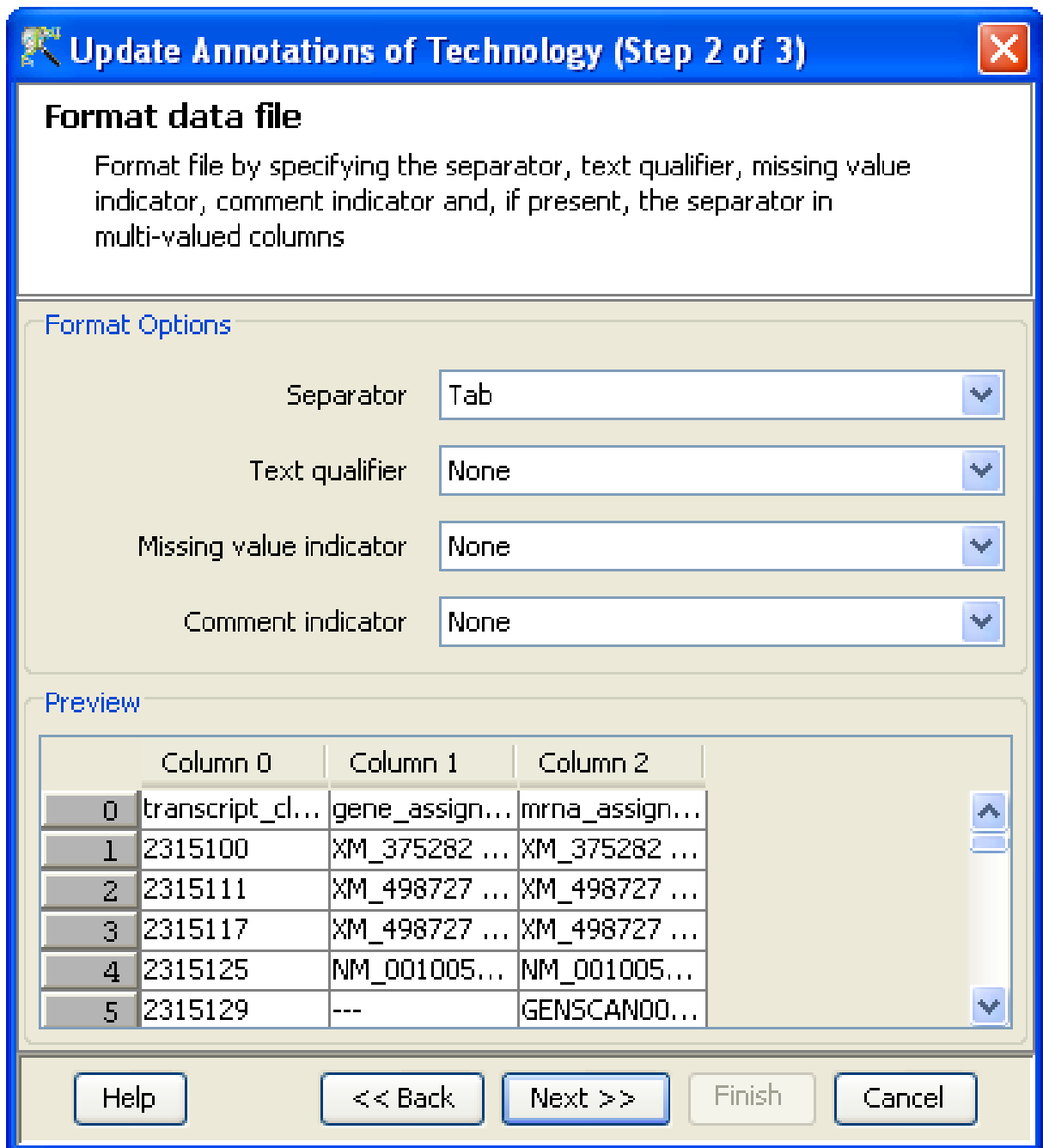


Figure 3.8: Format data file

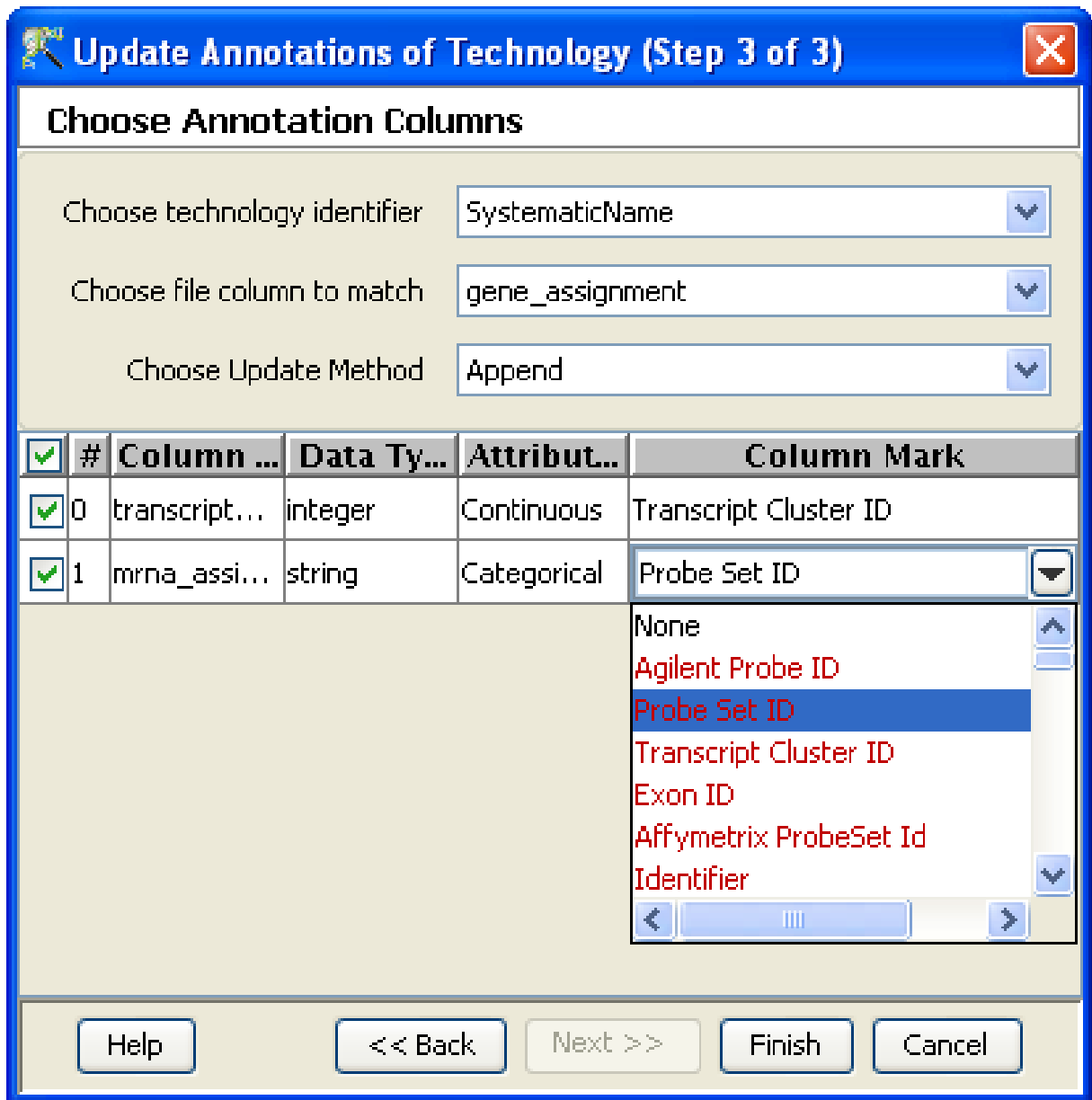


Figure 3.9: Choose Annotation Columns

homologs (similarity attributable to descent from a common ancestor) among the annotated genes of several completely sequenced eukaryotic genomes is performed using HomoloGene.

GeneSpring GX could have multiple experiments open at the same time. Exactly one of these experiments is active at any time. The desktop in the center shows views for the active experiment. You can switch active experiments by either clicking on the title bar of the experiment in the **Navigator**, or by clicking on the tab title of the experiment in the main **Desktop**. When the active experiment is changed, the active entity list of the project is also changed to the “All Entities” entity list of that experiment.

As mentioned before, if you click on another entity list of the active experiment, all views of that experiment are restricted to show only the entities in that entity list. In addition if you click on an entity list of an experiment other than the active one, the views are still constrained to show only that entity list.

Note that if the two experiments do not correspond to the same technology then entities in the entity list will need to be translated to entities in the active experiment. **GeneSpring GX** does this translation seamlessly for a whole range of organisms which are given in the table below.

Serial No.	Organism
1	Mus musculus
2	Rattus norvegicus
3	Magnaporthe grisea
4	Kluyveromyces lactis
5	Eremothecium gossypii
6	Arabidopsis thaliana
7	Oryza sativa
8	Schizosaccharomyces pombe
9	Saccharomyces cerevisiae
10	Neurospora crassa
11	Plasmodium falciparum
12	Caenorhabditis elegans
13	Anopheles gambiae
14	Drosophila melanogaster
15	Danio rerio
16	Pan troglodytes
17	Gallus gallus
18	Homo sapiens
19	Canis lupus familiaris
20	Bos taurus

Table 3.1: HomoloGene Table

This cross-organism translation is done via HomoloGene tables <ftp://ftp.ncbi.nih.gov/pub/HomoloGene> that map Entrez identifiers in one organism to Entrez identifiers in the other.

Consider a technology T1 from vendor V1 (Affymetrix, Illumina, Agilent,Generic(Entrez-ID must be present)) for organism O1 (Ex:Human) and another technology T2 from vendor V2 (Affymetrix, Illumina,

Agilent, Generic(Entrez-ID must be present)) for organism O2 (Ex:Rat)

Translation compares the two cases:

T1V1O1=T2V2O2 via Entrez ID in the following situations:

- between same organism but different technologies
- between different organisms and different technologies.

3.3.1 Implementation

Translation is performed using Entrez Gene ID. The identifiers of the entity list to be translated are used to get the corresponding Entrez gene IDs say for technology T1. Using Homologene data, Entrez gene IDs are then retrieved for technology T2. These are then mapped to the identifiers of T2.

How is translation done?:

There are two ways to perform translation. The first method involves the following steps:

- Consider Entity list En1 from an experiment in T1 to be translated to T2.
- Keeping E2 as the active experiment, click on En1 in E1.
- This will restrict the view in E2 to the entity list selected in E1.
- Using this view (Spreadsheet, Box whisker, Profile plot), go to toolbar icon (create entity list) and create the entity list En2.

Alternative method to do translation involves Right clicking on En1 on E1 and selecting the option **Translate list**.

- Step 1 of the Translation Inspector wizard appears. This is the **Input parameters** page and you can import values associated with your entity list such as p-value, fold change etc along with either the raw or normalized signal values. Also, Interpretation can be chosen here from the drop down. By default, 'All samples' is chosen. Click **Next** to proceed. See Figure 3.10
- Step 2 shows the **Translation Table** page which has two tabs, the **Translated List** and **Translation mapping**. The Translation mapping table shows the mapping of the original entity list to the destination technology along with the annotations. The Translated list shows a list of probe-sets(destination technology identifiers) which represent the number of entities that have been translated along with the list associated values of the original entity list. The following rules are applied to the data associated with the entity lists while performing Translation:

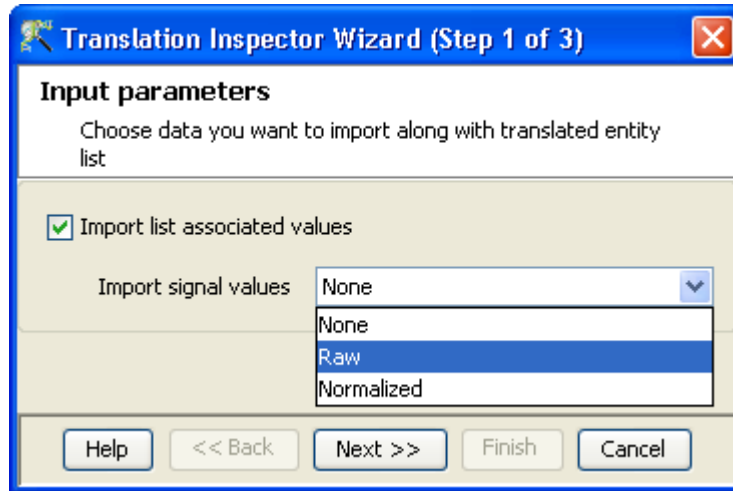


Figure 3.10: Input Parameters

- The first scenario is when multiple entities of the entity list correspond to one entity of the active dataset.

For example, when Translation is performed from Affymetrix HG_U95Av2 to Agilent Two-dye technology-12097, then values in the 'Translated List' would then correspond to the average of the 2 entities eg., the Agilent Probeset ID: A_23_P209059 corresponds to the Affymetrix probeset IDs: 38521_at and 38522_s_at. In the Translated List, values of the Affymetrix samples corresponding to the Probeset A_23_P209059 would be an average of 38521_at and 38522_s_at.

- The other scenario is when one entry of the entity list corresponds to multiple entries in the active data set.

If the above example is reversed, then the Probeset IDs A_23_P93015 and A_23_P85053 correspond to the Affymetrix 38523_f_at probeset. In this case, in the Translated List, the Probeset IDs A_23_P93015 and A_23_P85053 would report the same values, as that of 38523_f_at.

Annotations can be configured using *Configure Columns* button. See Figure 3.11

- Step 3 shows the **Save Entity List** window. This displays the details of the entity list created as a result of translation such as Creation date, modification date, owner, number of entities, notes etc. Click *Finish* and an entity list will be created and will be displayed in the experiment navigator of the destination experiment i.e., E2. Annotations can be configured using *Configure Columns* button. See Figure 3.12

Now any further analysis can be done and compared between En1 and En2.

3.3.2 Explicit Translation mapping

GeneSpring GX provides a way to explicitly define an annotation column for the source technology and an annotation column for the destination technology for translation, through the menu *Tool* → *Options*

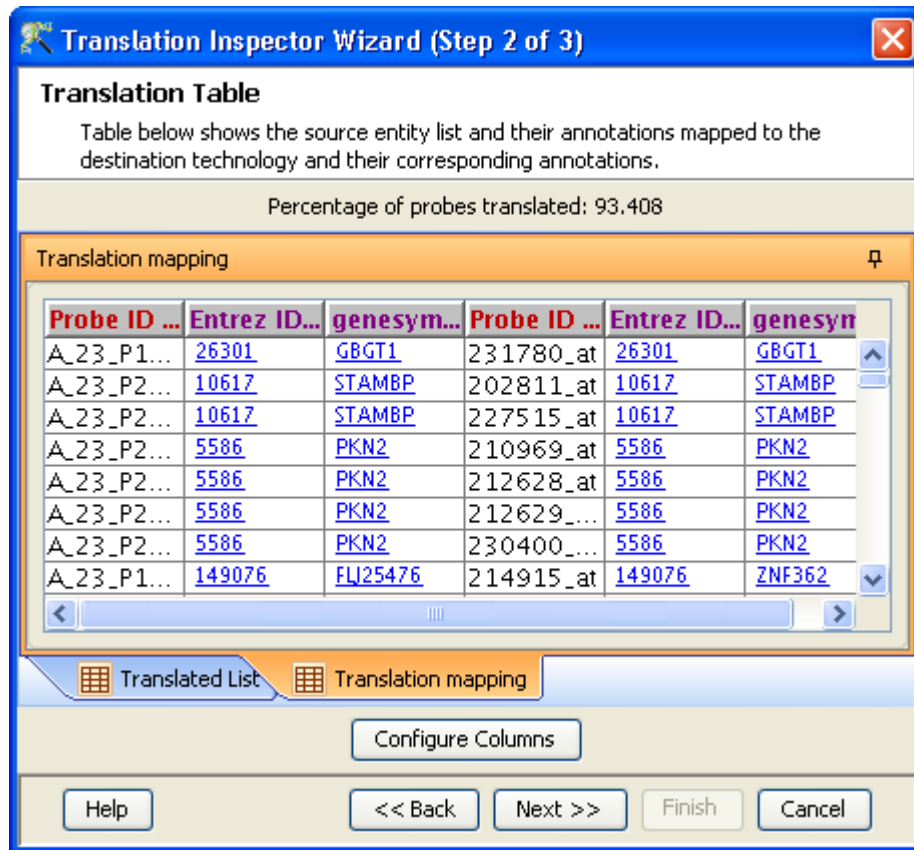


Figure 3.11: Translation Table

→ *Miscellaneous* → *Translation Mapping*. Note that this explicit mapping will override the default EntrezID mapping. This feature is useful in translating data between a custom technology and a standard technology.

Go to *Tool* → *Options* → *Miscellaneous* → *Translation Mapping*. The window will allow the user to define the source and destination technologies, along with the name of those columns. There is a provision to add or remove technologies. An Error messages is shown if the source and destination technology are the same. If a mapping is already defined, duplicate mapping will not be allowed.

A typical use case is that of handling Affy text files during migration from GX 7.0 to GX 11.0. Migration tool cannot understand the text files as those of Affymetrix technology and an explicit mapping achieves the translation effortlessly.

Note: Explicit translation mapping will override the default EntrezID mapping. Duplicate mapping will not be allowed.

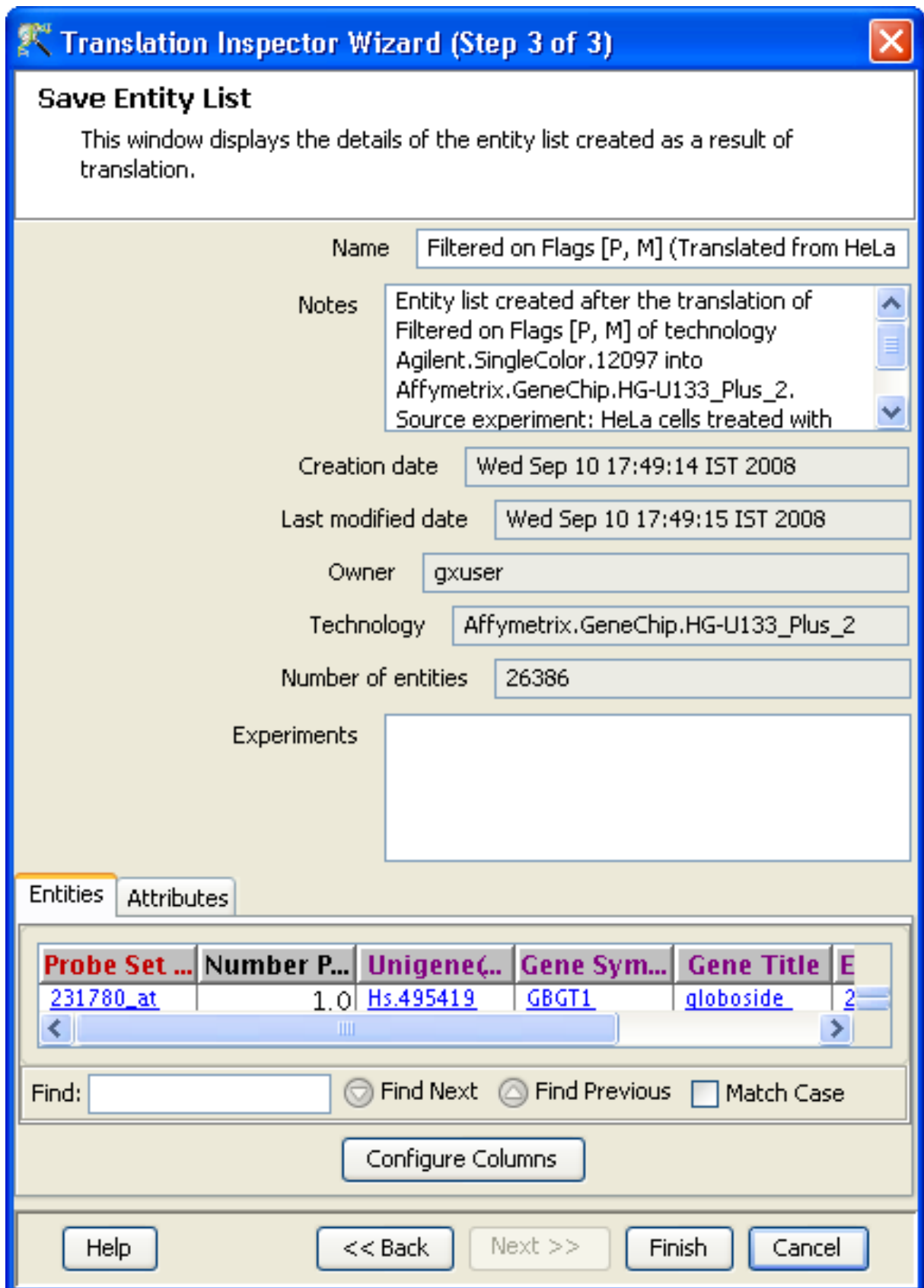


Figure 3.12: Save Entity List

3.3.3 Translation in Copy number and Association experiments

Translation in copy number and association experiments are slightly restricted in terms of the columns used as ID; see section [Entity Lists and Translation rules in copy number](#) for complete details. Note that explicit translation mapping does not work with copy number and association experiments.

3.4 Biological Genome

A Biological Genome refers to the collective set of all major annotations (Entrez-ID, GO IDs etc.) for any particular organism. It is created using the information available at NCBI and can be stored in **GeneSpring GX**. It is independent of any chip technology and once created can be used across multiple chip types and technologies. Biological Genome creation uses the following files from the NCBI site: All_Data.gene.info, gene2accession, gene2go, gene2refseq and gene2unigene. The NCBI site used for Biological Genome creation can be accessed from *Tools* → *Options* → *Miscellaneous* → *NCBI ftp URL*. Since the Standard Technologies available from the update server usually contain all the annotations, Biological Genome is useful mainly in cases of custom technologies.

Biological Genome is essential in performing biological analyses in Generic experiments lacking annotations. For eg., if a particular experiment does not have GO annotation columns, then the same can be obtained from Biological Genome and GO analysis can be performed. The Biological Genome can be created from *Annotations* → *Create Biological Genome* using the following steps:

- On selecting *Annotations* → *Create Biological Genome*, a window appears with a list of organisms for which biological genomes can be created. This allows the user to select the species of interest.
- The user is also presented with an option to download the genomic data either from the NCBI ftp site or from a local folder. See figure 3.13. If the option to download from the NCBI site is chosen, then a confirmation window appears.
- On choosing to go ahead, the user has to specify the folder in the system into which the files can be downloaded. This is a one time process as once the folder is created; subsequent creation of genomes for other organisms can be done from this folder by choosing the *Use from local folder* option. Alternatively the user can choose to download the files from the NCBI site directly into a local folder and utilize the option *Use from local folder* for the genomic data.

For using the Biological Genome created for an organism in an experiment, the user has to update the annotations for that particular technology from Tools-Update Technology Annotations-Update from Biological Genome. For more details on updating annotations, refer to [Update Technology Annotations](#).

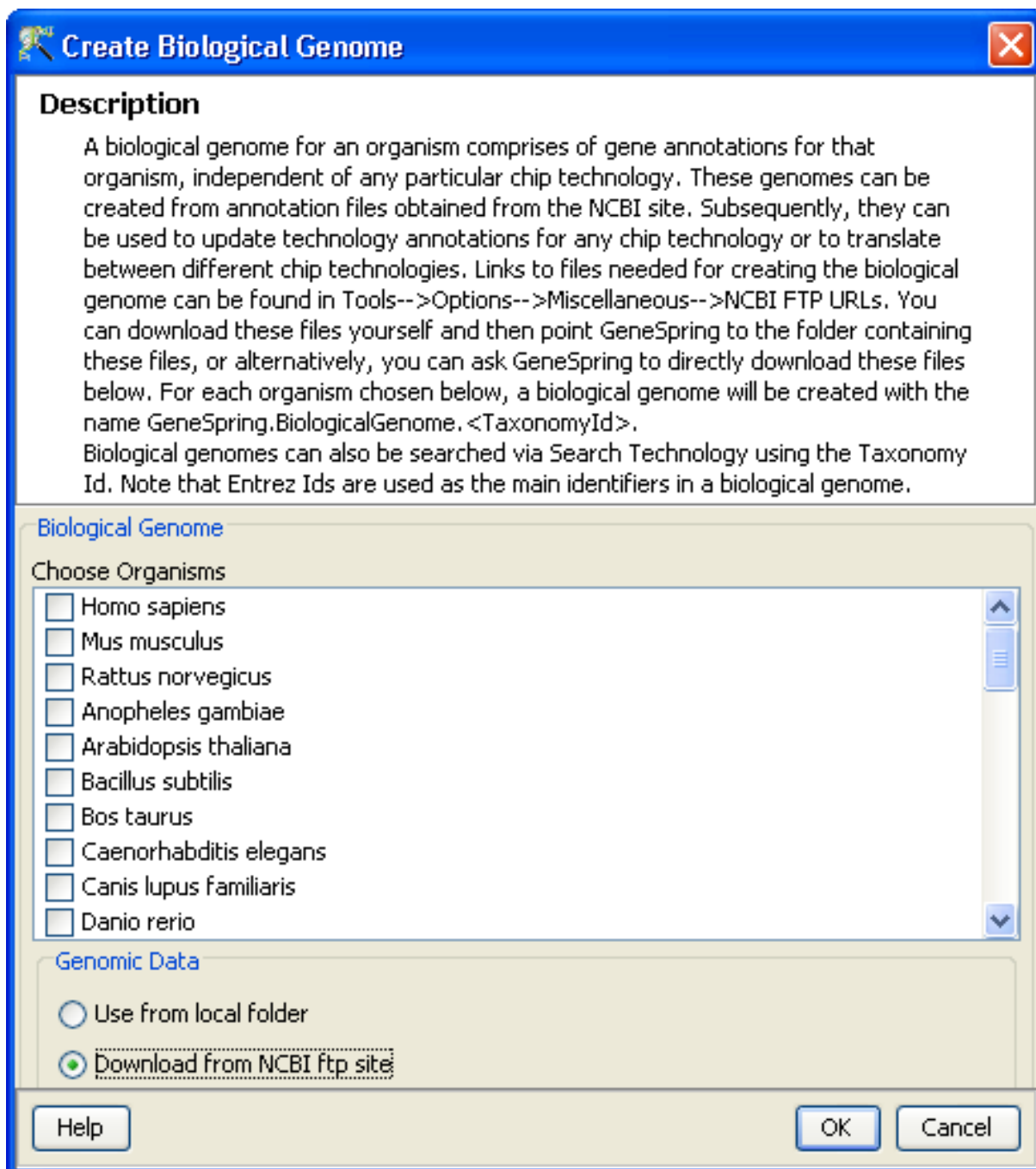


Figure 3.13: Create Biological Genome

Chapter 4

Data Migration

The following sections explain about all the various data migration processes in **GeneSpring GX** .

4.1 GeneSpring GX Data Migration from GeneSpring GX 7

Migration in **GeneSpring GX** happens genome by genome. Migration of a genome involves, migrating the corresponding samples, experiments, genelists, trees and also the hierarchy of the involved objects. From here on, the phrase migration of a genome implies migration of all the above objects. Migration of data from **GS7** to **GX11** involves the following steps.

4.1.1 Migrations Steps

Step 1 This step is needed only if **GS7** and **GX11** are installed on separate machines. In this case, copy the **Data** folder from **GS7** to any location on (or accessible from) the machine where **GX11** is installed. The **Data** folder for **GS7** is located inside its installation folder.

Step 2 Launch **GX11** now and run *Tools*→*Prepare for GS7 Migration*. Then provide the location of the **Data** folder described in Step 1 and click on the **Start** button. See Figure 4.1. This launches a procedure with the following properties:

- This procedure prepares the **Data** folder for migration to **GX11** . Note that this procedure does not itself perform migration.
- This is a one-time procedure. Once finished, you can migrate genomes from **GS7** to **GX11** using the steps described further below. If any new experiment is added after the entire migration of the genome is finished, one has to run step 2 again. However only new experiments added will be migrated. Any changes to the existing experiments will not be reflected in the already migrated experiments.

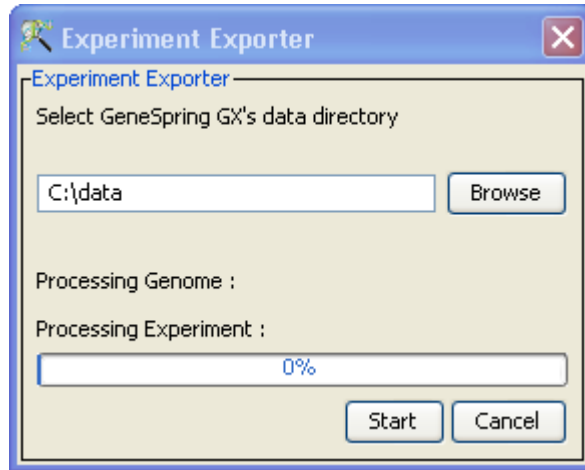


Figure 4.1: Experiment Exporter

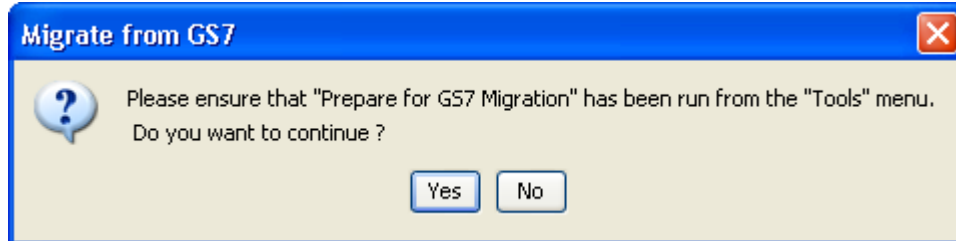


Figure 4.2: Confirmation Window

- This procedure could be time consuming; a typical run comprising 28 experiments takes about 20 minutes. You can reduce the time needed by running Step 2 only on specific genomes of interest. To do this, create a new folder called XYZ (anywhere), then simply copy the relevant genome subfolder from the **Data** folder to within XYZ. Finally, in the dialog for Step 2, provide XYZ instead of the **Data** folder.
- This procedure could give errors for two known reasons. The first situation is when it runs out of space in the system's temporary folders (on Windows systems this would typically be on the C: drive). If this happens then clear space and start Step 2 again. The second situation is when the **GS7** cache file encounters an internal error; this could result in Step 2 hanging. In this situation, delete the cache file inside the *Data* folder and restart Step 2.

Step 3 This step and subsequent steps focus on particular genome of interest. To migrate this genome from **GS7** to **GX11**, run *Tools* → *Migrate from GS7*. This will ask for the confirmation of the user whether Step 2 has been run on the genome. If Step 2 is not run, click **Cancel**. Note that genomes on which Step 2 hasn't been run will not be migrated. If Step 2 has been run on the genome of interest, click **Ok** and proceed further. See Figure 4.2.

Step 4 The **GS7 Data** folder needs to be provided at this step. Then **GX11** will automatically detect all **GS7** genomes within this **Data** folder. This will launch a window which shows the genome(s) selected for migration. By default all the genomes, which haven't been migrated before will be

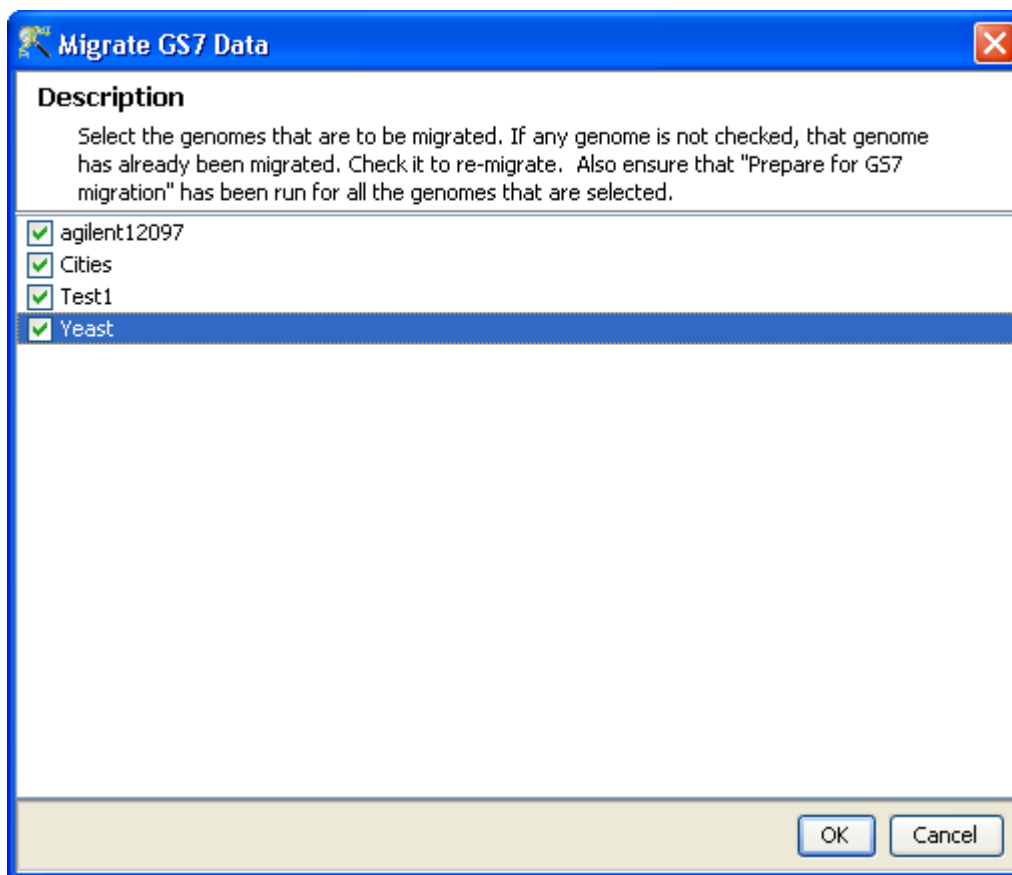


Figure 4.3: Migrate GS7 Data

selected. Select only the genome(s) to be migrated and click **OK**. See Figure 4.3. If the genome(s) was partially migrated before, it will launch another window showing the partially migrated genome(s). See Figure 4.4. Select the genome if a fresh migration has to be done or just click **Ok** to resume migration the genome from the point where it was left off. This step will eventually launch a progress bar showing the status of the migration. Migration can be a time consuming process depending on the amount of data to be migrated.

The General rate of migration is listed in the table below:

Task	Speed
Prepare for GS7 Migration	10 sample exp/1 min
Sample Migration	5MB/sec
Experiment Migration	()10 sample exp/1 min
Other Objects	4 Objects/sec

Table 4.1: Migration Rate

The Migration timings for a HG_U133_Plus2 genome are listed in the tables below:

Machine: Windows XP Genuine Intel P4, 2.9Ghz, 1GB RAM, Xmx set to 1024m

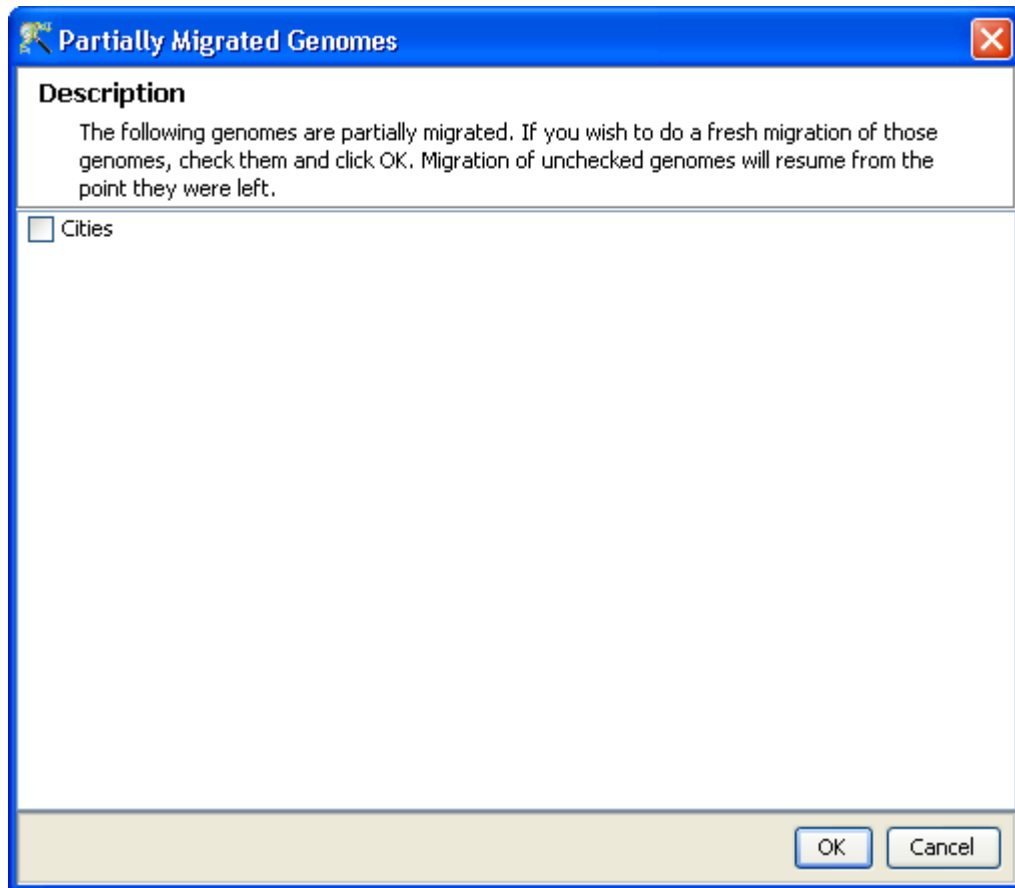


Figure 4.4: Partially Migrated Genomes

Task	Speed
Prepare for GS7 Migration	10 sample exp/1 min
Sample Migration	5MB/sec
Experiment Migration	500 sample exp/ 90 min

Table 4.2: Migration Rate on Windows OS

Machine: Debian OS, Intel Xeon CPU X3220 2.4Ghz Quad Core 32 bit, 2GB RAM, Xmx set to 1500m

Task	Speed
Prepare for GS7 Migration	10 sample exp/1 min
Sample Migration	5MB/sec
Experiment Migration	700 sample exp/ 65 min

Table 4.3: Migration Rate on Debian OS

To migrate experiments with around 1500 samples, the user needs to use a high end machine (64 bit,

8GB RAM)

The data that is brought in from **GS7** will undergo the following transformations:

- **GX11** works with data on the base 2 logarithmic scale while normalized values coming from **GS7** are in linear scale; these are therefore converted to the log scale in **GX11** .
- Prior to log transformation, **GX11** will threshold the data so all values below 0.01 are thresholded to 0.01; this is consistent with **GS7** as well.

4.1.2 Migrated Objects

When a **GS7** experiment is migrated to **GX11** , the following changes happen to objects contained therein.

All experiments other than Affymetrix and Agilent experiments with standard technologies will be migrated as what are called "custom" experiments. Each custom experiment will have Raw, Normalized and Control values exactly as derived from **GS7** , with just the following change: normalized values will be displayed on the log scale while Raw and Control values will be displayed on the linear scale, in entity inspectors.

For Affymetrix experiments with standard technologies, Raw and Normalized values will be migrated from **GS7** , with raw values kept in the linear scale and normalized values reported on the log scale. For Agilent single color experiments with standard technologies, normalized values will be migrated from **GS7** and raw values will be reread from the associated sample files (which may take some time). For Agilent two color experiments with standard technologies, normalized values will be migrated from **GS7** and raw Cy3 and Cy5 values will be reread from the associated sample files (which may take some time). In both cases, raw values are reported on the linear scale while normalized values are on the log scale.

Experimental Parameters and Interpretations: All experimental parameters, parameter values for each such parameter, and the order of these values for each such parameter are migrated. All interpretations are migrated as well. However the following things need to be noted.

GS7 and **GX11** use interpretations slightly differently. **GX11** does away with the notion of continuous/non-continuous etc causing profile plots launched on an interpretation to be slightly different. For instance, **GS7** considers non-continuous parameters first and continuous parameters later in creating a profile plot, while **GX11** considers parameters in the order in which they appear on the experimental grouping page. So if a profile plot in **GX11** for a particular interpretation feels different from the corresponding plot in **GS7** , try modifying the order of parameters and the order of parameter values on the experimental grouping page; very often this will result in a similar plot in **GX11** .

Other Objects: Other objects like bookmarks, pathways etc are not migrated.

A complete description of the migrated objects and their association with the experiments is described in the section below and holds good for both **GS7** to **GX11** and **WG5.2** to **WG11** migrations. However, users/groups and permissions/ownerships are not applicable for the former.

4.2 Data Migration from WG5.2 to WG11

This section describes how various data objects from the **WG5.2** server appear in the **WG11** server after server migration has been performed as described in the **GeneSpring Workgroup Server** documentation. The **GeneSpring Workgroup Server** documentation is reachable using your web browser via the following url (here GSWG_server_IP_address needs to be filled in with the **WG11** machine IP address).

`http://GSW_server_IP_address:8080/WorkgroupServer`

The key difference in data organization between **WG5.2** and **WG11** is that **WG11** has project-centric hierarchical organization while **WG5.2** had a genome-centric flat organization. The process of migration tries to closely maintain the **WG5.2** perception while introducing **WG11** organization.

The following objects are migrated; details of each of these appear in the sections below, in turn.

- Users, Groups
- Samples
- Genomes, Projects, Experiments.
- Entity Lists, Gene Trees, Condition Trees and Classifications
- Ownership and Permissions

4.2.1 Users and Groups

For each user on the **WG5.2** server, a corresponding user account on the **WG11** server is created. Passwords are not migrated: each user gets a preset password, namely username123. The administrator account in **WG5.2** maps to a corresponding administrator account in **WG11**. For each group of users in **WG5.2**, a corresponding group is created in **WG11**. In addition, one extra group called *Everyone* is created and all users are members of this group. Ownership and permissions for the various objects and the various users and groups will be described after these objects have been described in the sections below.

4.2.2 Samples

To describe this in more detail, we need to understand the various constituents of a sample in **WG5.2** .

- **Input Files:** A typical sample was imported into **WG5.2** starting with an input file, e.g., a CEL file or a .txt file. Typically, one input file contains one sample. However, there are exceptions; one input file could contain multiple samples, as in the case of an Illumina input file. And there is the rare case of Imagene generated two color raw files, where two input files together constitute one sample.
- **Processed Sample:** A processed sample is what is created from the above input files in **WG5.2** . This processed sample contains the relevant segment of data from the input file with further transformations.
- **Other Attachments:** An attachment is an auxiliary file associated with a sample in **WG5.2** . A sample could have one or more attachments, for instance, DAT files, ARR files etc.
- **Sample Attributes:** Attributes are other key value pairs associated with a sample and used typically for search.

When migrated into **WG11** , a sample from **WG5.2** has the following possible outcomes.

Migration to Standard Samples: A standard sample is one for which **GeneSpring GX 11.0** understands the file format off-the-shelf and new experiments can be created with such samples in **WG11** directly via the Create New Experiment wizards. In addition, input files for these samples can be downloaded from the experiment navigator by right-clicking on the sample. Most Affymetrix samples (except those based on custom CDFs) and Agilent samples obtained from FE versions 8.5.x and 9.5.x will be converted to standard samples.

Migration to Raw Samples: A raw sample is one for which **GeneSpring GX 11.0** does not understand the file format off-the-shelf. Other than Affymetrix samples (except those based on custom CDFs) and Agilent samples obtained from FE versions 8.5.x and 9.5.x, all others will be converted to raw samples. To create new experiments with such raw samples, one needs to follow a multi-step process. First, use the **GeneSpring Manager** to identify input raw files associated with these samples. To do this, log into the **GeneSpring Manager** , use the *Search* → *All* menu item and choose RawFiles as the object type. Then download one of the resulting raw files; these files will have the same name as the corresponding samples. Second, create a new *custom technology* from the downloaded raw file via *Tools* → *Create Custom Technology* in **GeneSpring GX 11.0** . And third, use that technology to create a new experiment from these raw files via *Create New Experiment* → *Custom* → *Choose Raw Files*. Note that there is an additional option for some Illumina multi-sample input files which **GeneSpring GX 11.0** recognizes; each of the above raw files will have this multisample file as an attachment viewable from the inspector in **GeneSpring Manager** ; download this multisample file and use the Create New Experiment function with this file to create a new Illumina experiment.

Migration Failure: This will only happen if the sample has no associated genome in **WG5.2** , or an associated genome that is faulty for some reason.

Migration Scheme:

For migration into **WG11** , each processed sample in **WG5.2** is considered in turn. For a particular processed sample, all its attachments in **WG5.2** are scanned to see if any of these represents an input file which **GeneSpring GX 11.0** can convert to a standard sample. If so, then that input file is migrated into **WG11** as a standard sample. All other attachments with the processed sample in **WG5.2** are added as attachments to this standard sample in **WG11** with the same ownership and permissions as the standard samples, and all attributes of the processed sample are made attributes of this standard sample. Otherwise, if none of its attachments in **WG5.2** represent an input file which **WG11** can convert to a standard sample, the processed sample is itself migrated into **WG11** as a raw sample (the associated technology name would be GS7.Custom.xxx), and the associated input files are migrated as raw files with the same corresponding sample names. All attachments/attributes with the processed sample are added as attachments/attributes to these corresponding raw files, with the same ownership and permissions as these raw files.

4.2.3 Genomes, Projects, Experiments

Objects in **WG5.2** were organized by genome, i.e., each object belonged to exactly one genome. In later versions of **WG5.2** , an extra project tag was introduced; objects tagged with a particular project tag could be viewed as one collection. In contrast, the organization in **WG11** is purely project based and not genome based. Hence the need for mapping from a genome based organization to a project based organization, which is done as follows.

For each genome in **WG5.2** , a special project called the *Genome Project* is created in **WG11** . This Genome Project contains all experiments associated with this genome. In addition, to reflect project tags on objects in **WG5.2** , special *Project Projects* are created in **WG11** ; a Project Project contains only those experiments which have the corresponding project tag in **WG5.2** .

Each of these experiments in turn contains other objects (Entity Lists, Gene Trees, Condition Trees and Classifications) associated with this genome. There are two cases here. If an experiment has a project tag then it contains only those objects which have the same project tag. And if an experiment has no project tag then it contains all objects in the genome which do not have any project tags; these objects appear classified into two groups, those which have no association with projects and those which do; the latter appear in appropriate folder structures which describe the project association. Within the above framework, the folder hierarchy for each object is preserved as in **WG5.2** .

The data in an experiment comprises normalized values, raw values and flags for each entity (gene) and each associated sample, and experimental grouping information. These are migrated directly from **WG5.2** , i.e., they are copied from **WG5.2** and not recalculated in **WG11** . Since algorithms and processing steps in **GeneSpring GX 7.3** and **GeneSpring GX 11.0** are different, further operations on this data could give slightly different results in **GeneSpring GX 7.3** and **GeneSpring GX 11.0** . For instance, if this experiment has samples that are CEL files then using these samples to resummaries and create a new experiment could give slightly different results.

4.2.4 Entity Lists, Gene Trees, Condition Trees and Classifications

The organization of objects (Entity Lists, Gene Trees, Condition Trees and Classifications) within an experiment is of course different in **GeneSpring GX 11.0** when compared to **GeneSpring GX 7.3** . All these objects appear within the Analysis subfolder in **GeneSpring GX 11.0** but retaining the same hierarchy as in **GeneSpring GX 7.3** . Rules for whether or not an object appears within a particular experiment are as in the paragraphs above.

4.2.5 Ownership and Permissions

Ownership in **WG11** is derived as follows:

All projects (Genome Projects and Project Projects) are set to be owned by the administrator. All other objects owned by a particular user are owned by the corresponding user in **WG11** . Objects owned by a group in **WG5.2** are also set to be owned by the administrator now (note **WG11** does not support the notion of group ownership).

Permissions for objects are derived as follows. If an object has read/write permissions for a particular user or group in **WG5.2** , the corresponding object has the same permissions for that user or group in **WG11** . There are two additional cases though. First, the owner of an experiment in **WG5.2** gets read and write permissions to both the Genome Project and the Project Projects (if any) which contain this experiment. Second, the members of a group which owns an object in **WG5.2** all get read/write permission to the corresponding object in **WG11** .

Another note for permissions in the context of objects stored in folder hierarchies. Consider an object O, say a gene tree, a condition tree or a classification, and suppose this object has permissions for a particular user. Further, suppose O appears nested inside one or more levels of folders in **WG5.2** and let F denote the parent folder (or any ancestor). If F does not have permissions for this user then O will not be visible to the user inside any of the relevant experiments; however, O will still be accessible via a search. On the other hand, if the user has permissions for F then the folder hierarchy above F and the object O will both be visible.

Finally a note on the administrator group. Non-administrator members of the administrator group do not automatically get access to objects owned by the administrator even though these objects are accessible to the administrator group. This is illustrated by the following example. Suppose user abc belongs to the administrator group which has say write access to an object O owned by the administrator. In **WG5.2** , abc will have write access to O. In **WG11** , abc will not have write access to O just by virtue of being part of the administrator group.

4.2.6 Potential causes of Migration failure and Known Issues

Some causes of migration failure and some known issues are listed below.

- Running out of RAM is one key issue; 8GB of RAM will ensure that experiments with up to 1500 HG_U133_Plus2 samples can be migrated.
- Unparseable characters in certain **GeneSpring GX 7.3** XML files, though rare, will cause the corresponding experiment to fail from being migrated.
- Enablement and Disablement of users is not migrated.
- Passwords are not migrated, instead new passwords of the form username123 are created.
- When **WG5.2** objects are migrated to **WG11** , the creation and modification dates of the new objects correspond to the date of migration rather than the date of creation/modification of the original **WG5.2** objects. The latter dates are added as user attributes (namely, **WG5.2** Creation Date and **WG5.2** Last Modification Date) and are available for search via the **GeneSpring Manager** .
- The administrator group is created but membership in this group is not migrated; so users will have to be added to this group explicitly.
- Condition trees that are malformed, possibly due to subsequent modification of conditions, may not be migrated.
- Occasionally, there might be experiments for which the corresponding genome is empty (possibly on account of a deletion event); such experiments will not be migrated.
- The administrator group in **WG5.2** behaves differently from the corresponding group in **WG11** . Suppose user abc belongs to the administrator group which has say write access to an object O owned by the administrator. In **WG5.2** , abc will have write access to O. In **WG11** , abc will not have write access to O just by virtue of being part of the administrator group.

4.3 Migration of GX11 Desktop Data to GX11 Workgroup

Migrating data from desktop to workgroup is a one time process. The tool migrates all the data on the local system to any user account on the workgroup in one shot. Data can only be migrated to a fresh user account, meaning, there should not be any data on the workgroup for that user. So before starting the migration the user has to ensure that he does not have any data on the workgroup. It is to be noted that migration once started CANNOT be aborted in between. The following steps need to be followed for Desktop to Workgroup migration:

1. Launch 'Migrate to Workgroup' from *Tools*→*Migrate to Workgroup*.
2. This launches a login dialog. The login details of the user to whom the data is to be migrated should be entered. Click *OK*.

3. If there are any custom technologies already existing on the Workgroup Server, this will launch a matching technologies dialog. To migrate these custom technologies with some other name, enter the appropriate name for every technology and click **OK**. If no change is made, that custom technology will NOT be migrated. All Standard technologies which exist on the Workgroup Server and also on the desktop will NOT be migrated. This step will start migrating all the data. This process cannot be aborted in between.

4.4 Migration of GeneSpring GX 10.0 to GeneSpring GX 11.0

Migration of **GeneSpring GX 10.0** experiments to **GeneSpring GX 11.0** happens when those experiments are opened in the updated **GeneSpring GX 11.0** product. This is done when the **GeneSpring GX 10.0** product is updated from *Help* → *Update Product*.

Chapter 5

Data Visualization

5.1 View

Multiple graphical visualizations of data and analysis results are core features of **GeneSpring GX** that help discover patterns in the data. All views are interactive and can be queried, linked together, configured, and printed or exported into various formats. The data views provided in **GeneSpring GX** are the [Spreadsheet](#), the [Scatter Plot](#), the [Profile Plot](#), the [Heat Map](#), the [Histogram](#), the [Matrix Plot](#), the [Summary Statistics](#), [Bar Chart](#), [MvA](#), [Genome Browser](#), [Plot List Associated Values](#) and the [Venn Diagram](#).

5.1.1 The View Framework in GeneSpring GX

In **GeneSpring GX** rich visualizations are used to present the results of algorithms. The user can interact with these views, change parameters and re-run the algorithm to get better results. The views also help in examining and inspecting the results and once the user is satisfied, these entity lists, condition trees, classification models, etc can be saved. The user can know the identity of a probe depicted by particular point on the view by pointing the mouse over it. You can also interact with the views and create custom lists from the results of algorithms. Details of the views associated with the guided workflow and the advanced workflow links will be detailed in the following sections.

In addition to presenting the results of algorithms as interactive views, views can also be launched on any entity list and interpretation available in the analysis from the view menu on the menu bar or from the tool bar. The [Spreadsheet](#), the [Scatter Plot](#), the [Profile Plot](#), the [Heat Map](#), the [Histogram](#), the [Matrix Plot](#), the [Summary Statistics](#), [Bar Chart](#), [MvA](#), [Genome Browser](#), [Plot List Associated Values](#) and the [Venn Diagram](#) view can be launched from the View menu on the menu bar. The views will be launched with the current active entity list and interpretation in the experiment.

Note: The key driving force for all views derived from the view menu are the current active interpretation and the current active entity list in the experiment. The conditions in the interpretation provide the columns or the axes for the views and the current active entity list determines the entities that are displayed as rows or points in the view. While making another entity list in the same experiment, the active entity list will dynamically display those entities in the current view. Clicking on an entity list in another experiment will translate the entities in that experiment to the entities in the current experiment (based upon the technology and the homologies) and dynamically display those entities.

5.1.2 View Operations

All data views and algorithm results share a common menu and a common set of operations. There are two types of views, the plot derived views, like the Scatter Plot, the Profile Plot, the Histogram and the Matrix Plot; and the table derived views like the spreadsheet, the heat map view, and various algorithm result views. Plot views share a common set of menus and operations and table views share a common set of operations and commands.

In addition, some views like the heat map are provided with a tool bar with icons that are specific to that particular data view. The following section below gives details of the common view menus and their operations. The operations specific to each data view are explained in the following sections.

Common Operations on Plot Views

See Figure 5.5

All data views and algorithm results that output a Plot share a common menu and a common set of operations. These operations are from Right-Click in the active canvas of the views. Views like the scatter plot, the 3D scatter plot, the profile plot, the histogram, the matrix plot, etc., share a common menu and common set of operations that are detailed below.

Selection Mode: All plots are by default launched in the Selection Mode. The selection toggles with the Zoom Mode where applicable. In the selection mode, left-clicking and dragging the mouse over the view draws a selection box and selects the elements in the box. Control + left-clicking and dragging the mouse over the view draws a selection box, toggles the elements in the box and adds to the selection. Thus if some elements in the selection box were selected, these would become selected and if some elements in the WQU3-2273-8247 selection box were unselected, they would be added to the already present selection.

Selection in all the views are lassoed. Thus selection on any view will be propagated to all other views.

Zoom Mode: Certain plots like the Scatter Plot and the Profile Plot allow you to zoom into specific portions of the plot. The zoom mode toggles with the selection mode. In the zoom mode, left-clicking and dragging the mouse over the view draws a zoom window with dotted lines and expands the box to the canvas of the plot.

Invert Selection: This will invert the current selection. If no elements are selected, Invert Selection will select all the elements in the current view.

Clear Selection: This will clear the current selection.

Limit to Selection: Left-clicking on this check box will limit the view to the current selection. Thus only the selected elements will be shown in the current view. If there are no elements selected, there will be no elements shown in the current view. Also, when Limit to Selection is applied to the view, there will be no selection color set and the elements will appear in the original color in the view. The status area in the tool will show the view as limited to selection along with the number of rows / columns displayed.

Reset Zoom: This will reset the zoom and show all elements on the canvas of the plot.

Copy View: This will copy the current view to the system clipboard. This can then be pasted into any appropriate application on the system, provided the other listens to the system clipboard.

Export Column to Dataset: Certain result views can export a column to the dataset. Whenever appropriate, the Export Column to dataset menu is activated. This will cause a column to be added to the current dataset.

Print: This will print the current active view to the system browser and will launch the default browser with the view along with the dataset name, the title of the view, with the legend and description. For certain views like the heat map, where the view is larger than the image shown, Print will pop up a dialog asking if you want to print the complete image. If you choose to print the complete image, the whole image will be printed to the default browser.

Export As: This will export the current view as an Image, an HTML file or the values as a text, if appropriate. See Figure 5.17

- **Export as Image:** This will pop-up a dialog to export the view as an image. This functionality allows the user to export a very high quality image. You can specify any size of the image, as well as the resolution of the image by specifying the required dots per inch (dpi) for the image. Images can be exported in various formats. Currently supported formats include png, jpg, jpeg, bmp or tiff. Finally, images of very large size and resolution can be printed in the tiff format. Very large images will be broken down into tiles and recombined after all the images pieces are written out. This ensures that memory is not built up in writing large images. If the pieces cannot be recombined, the individual pieces are written out and reported to the user. However, tiff files of any size can be recombined and written out with compression. The default dots per inch is set to 300 dpi and the default size of individual pieces for large images is set to 4 MB and tiff image without tiling enabled. These default parameters can be changed in the tools → Options dialog under the **Export as Image**. See Figure 20.7 and Figure 5.3

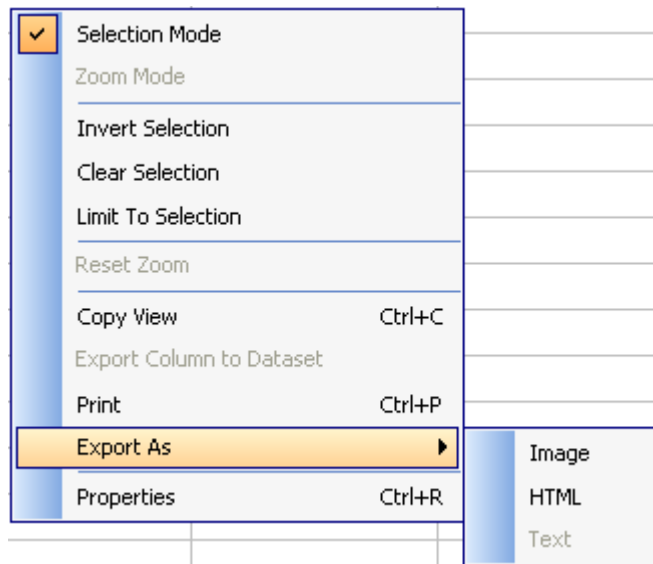


Figure 5.1: Export submenus

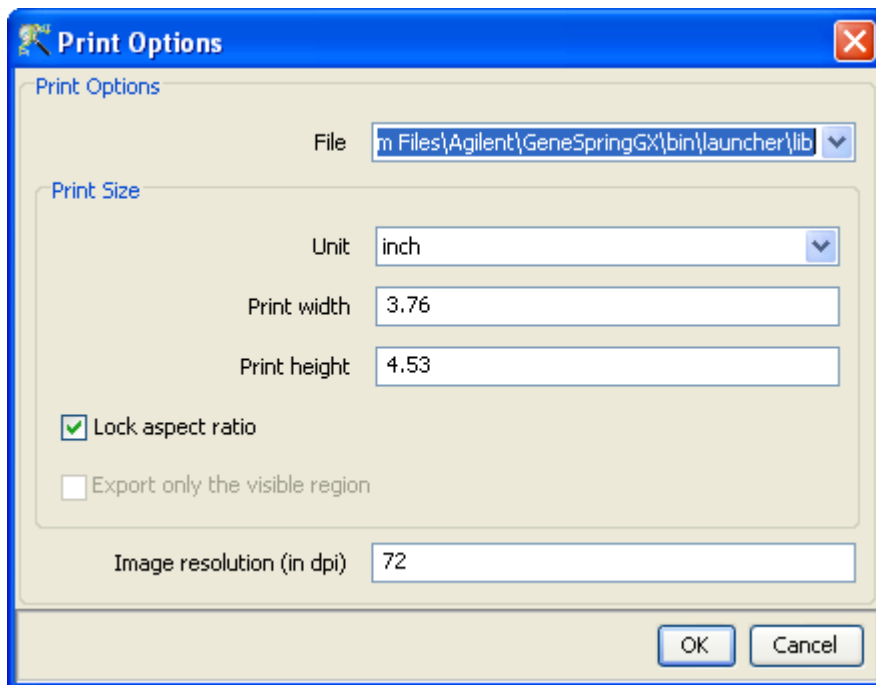


Figure 5.2: Export Image Dialog

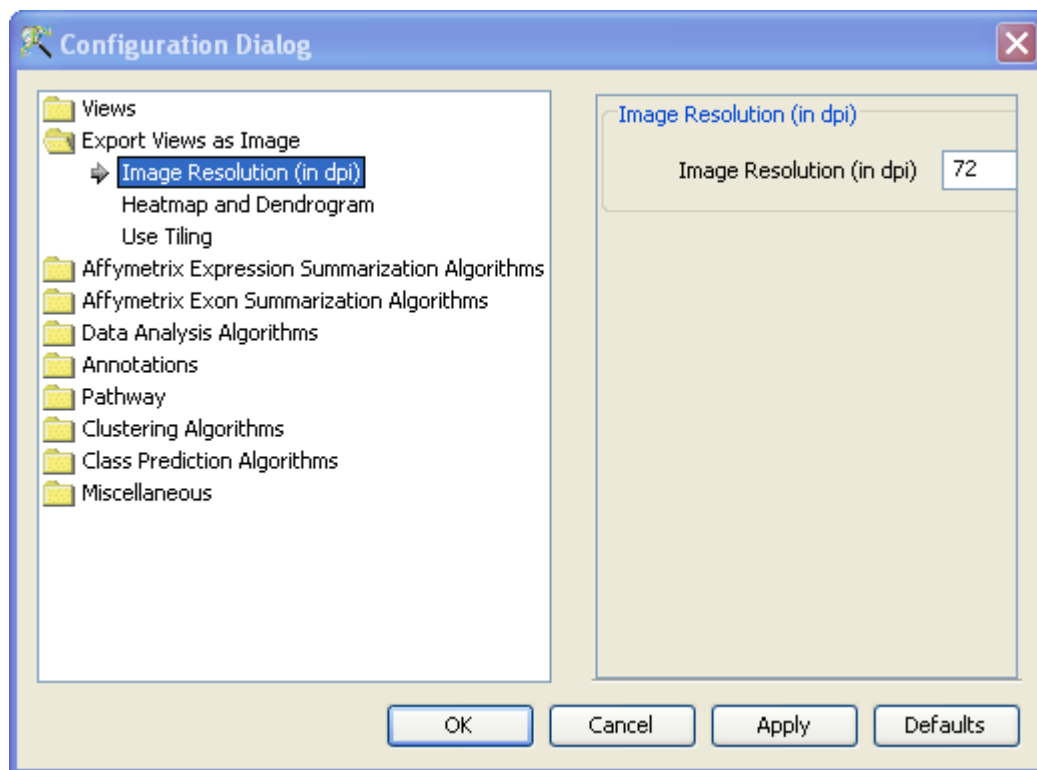


Figure 5.3: Tools → Options Dialog for Export as Image

Note: This functionality allows the user to create images of any size and with any resolution. This produces high-quality images and can be used for publications and posters. If you want to print vary large images or images of very high-quality the size of the image will become very large and will require huge resources. If enough resources are not available, an error and resolution dialog will pop up, saying the image is too large to be printed and suggesting you to try the tiff option, reduce the size of image or resolution of image, or to increase the memory available to the tool by changing the `-Xmx` option in `INSTALL_DIR/bin/packages/properties.txt` file. On **Mac OS X** the Java heap size parameters are set in in the file `Info.plist` located in `INSTALL_DIR/GeneSpringGX.app/Contents/Info.plist`. Change the `Xmx` parameter appropriately. Note that in the Java heap size limit on Mac OS X is about 2048M. See Figure 20.8

- **Export as HTML:** This will export the view as a html file. Specify the file name and the the view will be exported as a HTML file that can be viewed in a browser and deployed on the web.
- **Export as Text:** Not valid for Plots and will be disabled.

'Export As' will pop up a file chooser for the file name and export the view to the file. Images can be exported as a jpeg, jpg or png and 'Export As Text' can be saved as txt file.

Properties: This will launch the Properties dialog of the view as limited to selection along with the number of rows / columns displayed. the current active view. All Properties of the view can be

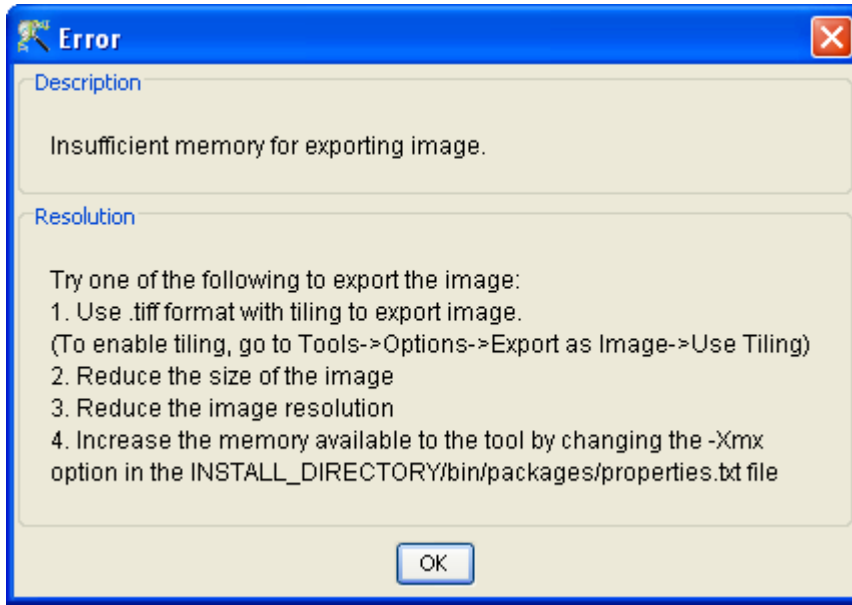


Figure 5.4: Error Dialog on Image Export

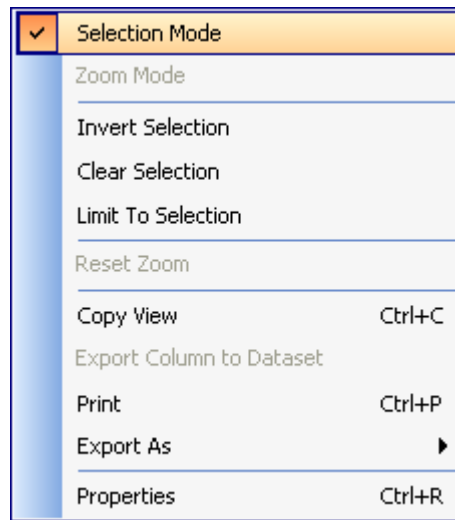


Figure 5.5: Menu accessible by Right-Click on the plot views

configured from this dialog.

Common Operations on Table Views

See Figure 5.6

All data views and algorithm results that output a Table share a common menu and a common set of operations. These operations are accessed from Right-Click in the active canvas of the views. Table views like Spreadsheet, the heat map, the Bar Chart, etc., share a common menu and a common set of operations that are detailed below.

Selection: The table views are by default launched in the Selection Mode. Either columns or rows or both can be selected on the Table. Selection on all views is lassoed. Thus selection on the table will be propagated to all other views of the data. All Table views allow row and column selection.

Clicking on a cell in the table will select the column or row or both column and row of the table. If clicking on a cell selects rows, Left-Click and drag the mouse. This will select all the rows. To select a large amount of continuous rows. Left-Click on the first row. Then scroll to the last row to be selected and Shift-Left-Click on the row. All rows between the first row and the last row will be selected and lassoed. Ctrl-Left-Click toggles the selection and adds to the current selection. Thus Ctrl-Left-Click on selected rows will unselect it, and Ctrl-Left-Click on unselected rows will add these rows to the selection.

Invert Row Selection: This will invert the current row selection. If no rows are selected, Invert Row Selection will select all the rows in the current table view.

Clear Row Selection: This will clear the current selection.

Limit to Selection: Left-Click on this check box will limit the table view to the current selection. Thus only the selected rows will be shown in the current table. If there are no selected rows, there will be no rows shown in the current table view. Also, when Limit to Selection is applied to the table view, there will be no selection color set and the rows will appear in the original color in the table view.

Select Column: This is a utility to select columns in any table view. Clicking on this will launch the Column Selector. To select columns in the table view, select the highlight the appropriate columns, move them to the Selected Items list box and click OK. This will select the columns in the table and lasso the columns in all the appropriate views.

Invert Column Selection: This will invert the current column selection. If no columns are selected, Invert Column Selection will select all the columns in the current table view.

Clear Column Selection: This will clear the current selection.

Copy Selected Columns: If there are any selected columns in the table, this will option will be enabled. Choosing this menu option will copy the selected column(s) on to the system clipboard. After copying to the clipboard, it will prompt an information messages saying it has Copied n column(s) to the clipboard. This can be later pasted into application on the user's desktop.

Copy View: This will copy the current view to the system clipboard. This can then be pasted into any appropriate application on the system, provided the other listens to the system clipboard.

Print: This will print the current active view to the system browser and will launch the default browser with the view along with the dataset name, the title of the view, with the legend and description. For certain views like the heat map, where the view is larger than the image shown, Print will pop up a dialog asking if you want to print the complete image. If you choose to print the complete image, the whole image will be printed to the default browser.

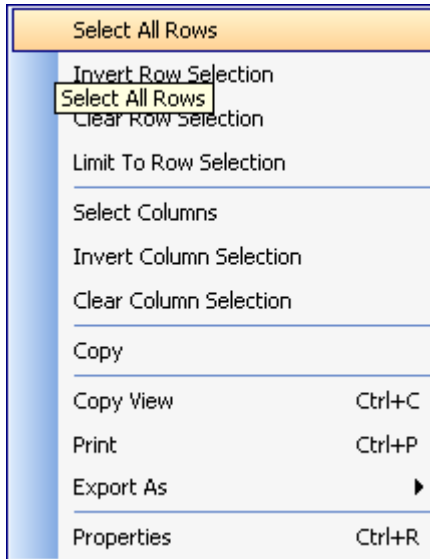



Figure 5.6: Menu accessible by Right-Click on the table views

Export As: This will the current view an Image, a HTML or as text. Export As will pop up a file chooser for the file name and export the view to the file. Images can be exported as a jpeg, jpg or png and Export as text can be saved as txt file.

Properties: This will launch the Properties dialog of the current active view. All Properties of the view can be configured from this dialog.

5.2 The Spreadsheet View

A spreadsheet presents a tabular view of the data. The spreadsheet is launched from the view menu with the active interpretation and the active entity list. Alternately, Left-Click on the tool bar 'Spreadsheet'  icon will launch the spreadsheet. The spreadsheet will display the normalized signal values of the conditions in the current active interpretation as columns in the table. If the interpretation is averaged, it will show the normalized signal values averaged over the samples in the condition.

The rows of the table correspond to the entities in the current active interpretation. Clicking on another entity list in the analysis tree will make that entity list active and table will be dynamically updated with the corresponding entity list.

Thus if the current active interpretation in an experiment is a time averaged interpretation, where the normalized signal values for the samples are averaged for each time point, the columns in the table will correspond to these averaged normalized signal values at each time condition. The rows of the table will correspond to the active entity list. In addition, the identifier for the entity and the default set of entity

ProbeName	US22502...	US22502...	US22502...	US225...
A_23_P146576	-0.27027...	0.3764472	-0.11248...	2.2192
A_23_P125016	-0.02642...	-0.10367...	0.299375...	0.0264
A_23_P28555	0.058861...	0.041716...	0.015798...	-0.015
A_23_P23227	0.2754221	0.123822...	0.325972...	-0.358
A_23_P137543	0.134859...	0.147473...	0.1287508	-0.147
A_23_P501193	-0.63219...	-0.54258...	-0.66686...	1.1783
A_23_P27247	0.194866...	0.4341488	-0.75057...	0.4126
A_23_P323270	-0.47815...	0.188868...	0.469717...	2.8949
A_23_P258433	-0.08102...	-0.07677...	0.076771...	-0.083
A_23_P28649	0.092969...	0.059925...	0.106556...	-0.071
A_23_P96529	-0.08102...	-0.07677...	0.076771...	-0.083
A_23_P2322	0.109666...	-0.17284...	0.213976...	0.0069
A_23_P372910	0.4508543	0.032173...	0.253333...	-0.354
A_23_P343357	-0.08102...	-0.07677...	0.076771...	-0.083
A_23_P501634	-0.11053...	0.027170...	0.8542857	-0.027

Figure 5.7: Spreadsheet

annotation columns will be shown. The legend window shows the interpretation on which the scatter plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the table will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the table. See Figure 5.7

5.2.1 Spreadsheet Operations

Spreadsheet operations are available by Right-Click on the canvas of the spreadsheet. Operations that are common to all views are detailed in the section [Common Operations on Table Views](#) above. In addition, some of the spreadsheet specific operations and the spreadsheet properties are explained below:

Sort: The Spreadsheet can be used to view the sorted order of data with respect to a chosen column. Click on the column header to sort the data based on values in that column. Mouse clicks on the column header of the spreadsheet will cycle through an ascending values sort, a descending values sort and a reset sort. The column header of the sorted column will also be marked with the appropriate icon.

Thus to sort a column in the ascending, click on the column header. This will sort all rows of the spreadsheet based on the values in the chosen column. Also an icon on the column header will

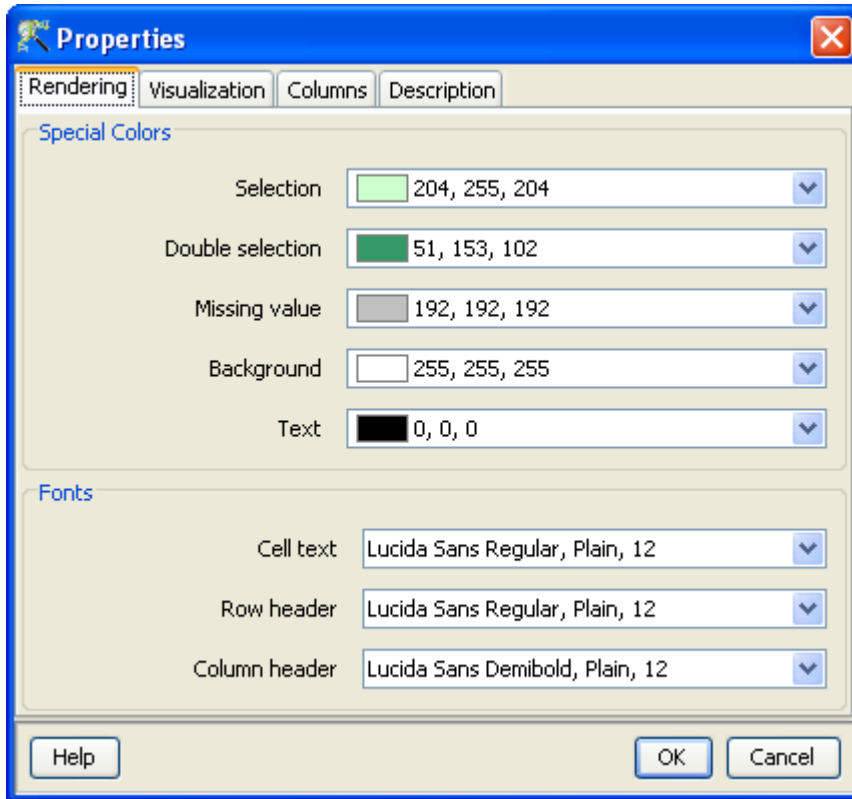



Figure 5.8: Spreadsheet Properties Dialog

denote that this is the sorted column. To sort in the descending order, click again on the same column header. This will sort all the rows of the spreadsheet based on the decreasing values in this column. To reset the sort, click again on the same column. This will reset the sort and the sort icon will disappear from the column header.

Selection: The spreadsheet can be used to select entities, and conditions Entities can be selected by clicking on any cell in the table. Conditions can be selected from the properties dialog of the spreadsheet as detailed below. The selection will be shown by the default selection color on the spreadsheet.

Entity Selection: Entities can be selected by left-clicking on any cell and dragging along the rows. Ctrl-Left-Click selects subsequent entities and Shift-Left-Click selects a consecutive set of entities. The selected entities can be used to create a new entity list by left-clicking on 'Create entity list from Selection'  icon. This will launch an entity list inspector where you can provide a name for the entity list, add notes and choose the columns for the entity list. This newly created entity list from the selection will be added to the analysis tree in the navigator.

5.2.2 Spreadsheet Properties

The Spreadsheet Properties Dialog is accessible by right-clicking on the spreadsheet and choosing **Properties** from the menu. The spreadsheet view can be customized and configured from the spreadsheet properties. See Figure 5.8

Rendering: The rendering tab of the spreadsheet dialog allows you to configure and customize the fonts and colors that appear in the spreadsheet view.

Special Colors: All the colors in the Table can be modified and configured. You can change the Selection color, the Double Selection color, Missing Value cell color and the Background color in the table view. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the properties dialog. To change a color, click on the appropriate color bar. This will pop-up a Color Chooser. Select the desired color and click OK. This will change the corresponding color in the Table.

Fonts: Fonts that occur in the table can be formatted and configured. You can set the fonts for Cell text, row Header and Column Header. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a Font, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Visualization: The display precision of decimal values in columns, the row height and the missing value text, and the facility to enable and disable sort are configured and customized by options in this tab.

The visualization of the display precision of the numeric data in the table, the table cell size and the text for missing value can be configured. To change these, Right-Click on the table view and open the Properties dialog. Click on the visualization tab. This will open the Visualization panel.

To change the numeric precision. Click on the drop-down box and choose the desired precision. For decimal data columns, you can choose between full precision and one to four decimal places, or representation in scientific notation. By default, full precision is displayed.

You can set the row height of the table, by entering a integer value in the text box and pressing Enter. This will change the row height in the table. By default the row height is set to 16.

You can enter any a text to show missing values. All missing values in the table will be represented by the entered value and missing values can be easily identified. By default all the missing value text is set to an empty string.

You can also enable and disable sorting on any column of the table by checking or unchecking the check box provided. By default, sort is enabled in the table. To sort the table on any column, click on the column header. This will sort the all rows of the table based on the values in the sort column. This will also mark the sorted column with an icon to denote the sorted column. The first click on the column header will sort the column in the ascending order, the second click on the column header will sort the column in the descending order, and clicking the sorted column the third time will reset the sort.

Columns: The order of the columns in the spreadsheet can be changed by changing the order in the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the *Selected items* to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the *Selected items* list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.

To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.

The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.3 MvA plot


The MvA plot is a scatter plot of the difference vs. the average of probe measurements between two samples. This plot is specifically used to assess quality and relation between samples. The MvA plot is used more in the two-color spotted arrays to assess the relation between the Cy3 and the Cy5 channels of each hybridization.

The MvA plot is launched from the view menu on the main menu bar with the active entity list in the experiment. Launching the plot from the menu in a two color experiment asks for the channel which can either be a sample or a condition depending on the interpretation chosen. It then shows the relation between the Cy3 and Cy5 channels of individual samples if the interpretation chosen is **All Samples**. In the case of other interpretations, it takes the average of Cy3 and Cy5 for all samples of a condition of the chosen interpretation to show the relation. See figure 5.9 In single color experiments, the plot asks for two inputs for calculating M and A. Depending on the interpretation chosen, the inputs could either be individual samples or conditions. The points in the MvA plot correspond to the entities in the active entity list.

Clicking on another entity list in the experiment will make that entity list active and the MvA plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the scatter plot.

The MvA Plot is a lassoed view, and supports both selection and zoom modes. Most elements of the MvA Plot, like color, shape, size of points etc. are configurable from the properties menu described in the properties section of [scatter plot](#).

5.4 The Scatter Plot

The Scatter Plot is launched from view menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar 'Scatter Plot'  icon will bring up the scatter plot. The Scatter Plot shows a 2-D scatter of all entities of the active entity list along the first two conditions of the active interpretation by default. If the active interpretation is an unaveraged interpretation, the axes of the scatter plot will be the normalized signal values of the first two samples. If the interpretation is averaged, the axes of the scatter plot will be the averaged normalized signal values of the samples in each condition. The axes of the scatter plot can be changed from the axes chooser on the view. The points in the scatter plot are colored by the normalized signal values of the first sample (or the averaged normalized signal values of the first condition) and are shown in the scatter plot legend window. The legend window also displays the interpretation on which the scatter plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the scatter plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the scatter

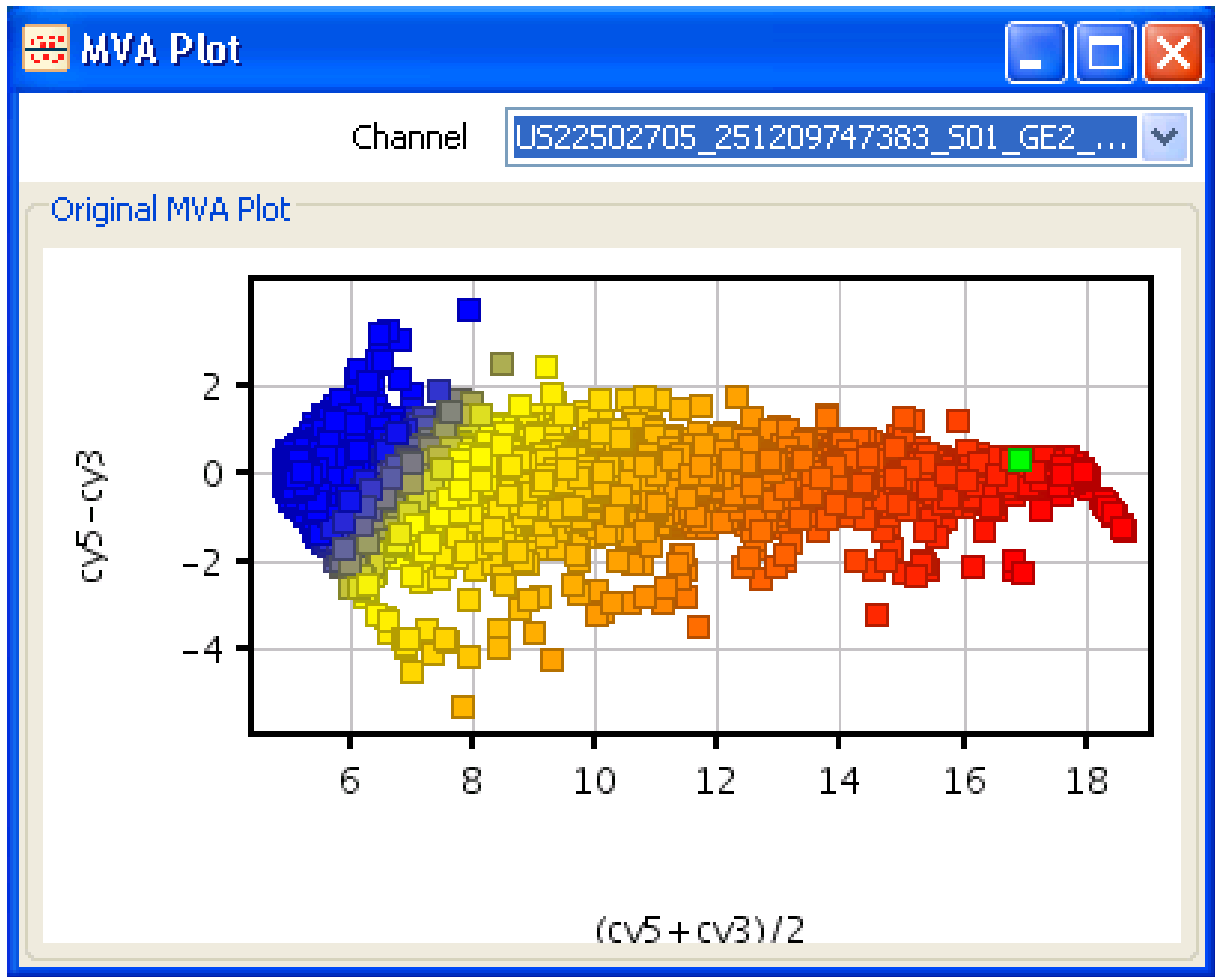


Figure 5.9: MvA plot

plot.

The Scatter Plot is a lassoed view, and supports both selection and zoom modes. Most elements of the Scatter Plot, like color, shape, size of points etc. are configurable from the properties menu described below. See Figure 5.10

5.4.1 Scatter Plot Operations

Scatter Plot operations are accessed by right-clicking on the canvas of the Scatter Plot. Operations that are common to all views are detailed in the section [Common Operations on Plot Views](#). Scatter Plot specific operations and properties are discussed below.

Selection Mode: The Scatter Plot is launched in the selection mode by default. In selection mode, Left-Click and dragging the mouse over the Scatter Plot draws a selection box and all entities within the

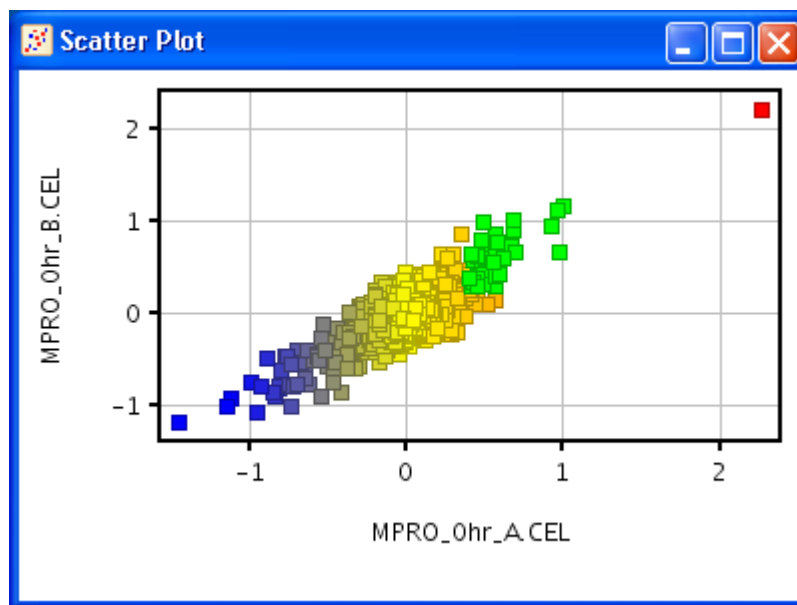



Figure 5.10: Scatter Plot

selection box will be selected. To select additional entities, Ctrl-Left-Click and drag the mouse over desired region. You can also draw and select regions within arbitrary shapes using Shift-Left-Click and then dragging the mouse to get the desired shape.

Selections can be inverted from the pop-up menu on Right-Click inside the Scatter Plot. This selects all unselected points and unselect the selected entities on the scatter plot. To clear the selection, use the Clear selection option from the Right-Click pop-up menu.

The selected entities can be used to create a new entity list by left-clicking on 'Create entity list from Selection'  icon. This will launch an entity list inspector where you can provide a name for the entity list, add notes and choose the columns for the entity list. This newly created entity list from the selection will be added to the analysis tree in the navigator.

Zoom Mode: The Scatter Plot can be toggled from the Selection Mode to the Zoom Mode from the right-click drop-down menu on the scatter plot. While in the zoom mode, left-clicking and dragging the mouse over the selected region draws a zoom box and will zoom into the region. Reset zoom from the right-click menu on the scatter plot, to revert back to the default, showing all the points in the dataset.

Save Entities: This allows the user to save entities with respect to the fold change. On selecting this option, a window appears with 4 boxes, each representing a set of entities satisfying a particular condition of fold change.

Scatter plot in Log10/Linear Scale: In **GeneSpring GX**, the data is usually in log 2 scale and the plots are launched with this data. In Scatter plot, there is an option to launch with data in log10 or linear scale, from menu *View* → *Plot Log10/Linear Values*. Refer section [Plot Log10/Linear Values](#) for details.

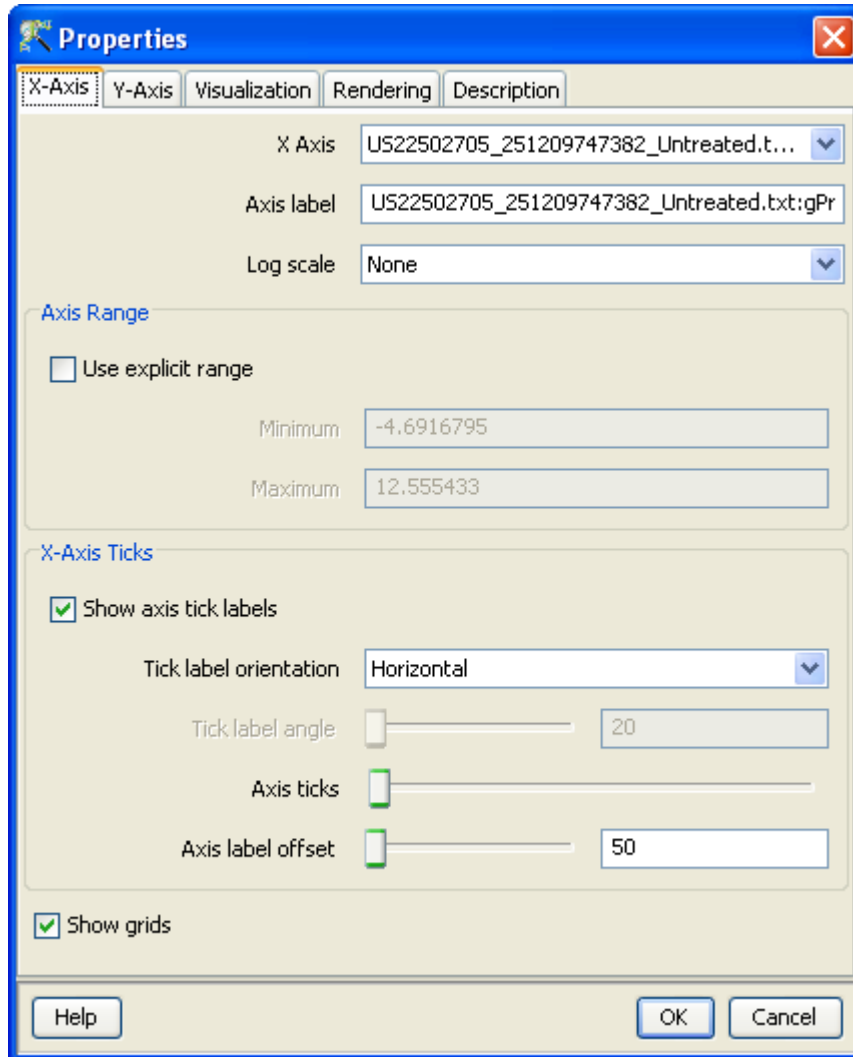


Figure 5.11: Scatter Plot Properties

5.4.2 Scatter Plot Properties

The Scatter Plot view offers a wide variety of customization with log and linear scale, colors, shapes, sizes, drawing orders, error bars, line connections, titles and descriptions from the Properties dialog. These customizations appear in three different tabs on the Properties window, labelled Axis, Visualization, Rendering, Description. See Figure 5.11

Axis: The axes of the Scatter Plot can be set from the Properties Dialog or from the Scatter Plot itself. When the Scatter Plot is launched, it is drawn with the first two conditions of the interpretation. These axes can be changed from the Axis selector in the drop down box in this dialog or in the Scatter Plot itself.

The axis for the plot, axis titles, the axis scale, the axis range, the axis ticks, tick labels, orientation

and offset, and the grid options of the plot can be changed and modified from the axis tabs of the scatter plot properties dialog.

To change the scale of the plot to the log scale, click on the log scale option for each axis. This will provide a drop-down of the log scale options.

None: If None is chosen, the points on the chosen axis is drawn on the linear scale

Log: If Log Scale is chosen, the points on the chosen axis is drawn on the log scale, with log of negative values if any being marked at missing values and dropped from the plot.

$$(if x > 0), x = \log(x)$$

$$(if x \leq 0), x = \text{missing value}$$

Symmetric Log: If Symmetric Log is chosen, the points along the chosen axis are transformed such that for negative values, the log of the 1- absolute value is taken and plotted on the negative scale and for positive values the log of 1+ absolute value is taken and plotted on the positive scale.

$$(if x \geq 0), x = \log(1 + x)$$

$$(if x < 0), x = -\log(1 - x)$$

To use an explicit range for the scatter plot, check this option and set the minimum and maximum range. By default, the minimum and maximum will be set to the minimum and maximum of the corresponding axis or column of the dataset. If explicit range is explicitly set in the properties dialog, this will be maintained even if the axis columns are changed.

The grids, axes labels, and the axis ticks of the plots can be configured and modified. To modify these, Right-Click on the view, and open the Properties dialog. Click on the Axis tab. This will open the axis dialog.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For continuous data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Visualization: The colors, shapes and sizes of points in the Scatter Plot are configurable.

Color By: The points in the Scatter Plot can be plotted in a fixed color by clicking on the Fixed radio button. The color can also be determined by values in one of the columns by clicking the 'By Columns' radio button and choosing the column to color by, as one of the columns in the dataset. This colors the points based on the values in the chosen columns. The color range can be modified by clicking the Customize button.

Shape By: The shape of the points on the scatter plot can be drawn with a fixed shape or be based on values in any categorical column of the active dataset. To change the 'Shape By' column, click on the drop down list provided and choose any column. Note that only categorical columns

in the active dataset will be shown list. To customize the shapes, click on the customize button next to the drop down list and choose appropriate shapes.

Size By: The size of points in the scatter plot can be drawn with a fixed shape, or can be drawn based upon the values in any column of the active dataset. To change the 'Size By' column, click on the drop down box and choose an appropriate column. This will change the plot sizes depending on the values in the particular column. You can also customize the sizes of points in the plot, by clicking on the customize button. This will pop up a dialog where the sizes can be set.

Drawing Order: In a Scatter Plot with several points, multiple points may overlap causing only the last in the drawing order to be fully visible. You can control the drawing order of points by specifying a column name. Points will be sorted in increasing order of value in this column and drawn in that order. This column can be categorical or continuous. If this column is numeric and you wish to draw in decreasing order instead of increasing, simply scale this column by -1 using the scale operation and use this column for the drawing order.

Error Bars: When visualizing profiles using the scatter plot, you can also add upper and lower error bars to each point. The length of the upper error bar for a point is determined by its value in a specified column, and likewise for the lower error bar.

If error columns are available in the current dataset, this can enable viewing Standard Error of Means via error bars on the scatter plot.

Jitter: If the points on the scatter plot are too close to each other, or are actually on top of each other, then it is not possible to view the density of points in any portion of the plot. To enable visualizing the density of plots, the jitter function is helpful. The jitter function will perturb all points on the scatter plot within a specified range, randomly, and then draw the points. The Add jitter slider specifies the range for the jitter. By default there is no jitter in the plots and the jitter range is set to zero. The jitter range can be increased by moving the slider to the right. This will increase the jitter range and the points will now be randomly perturbed from their original values, within this range.

Connect Points: Points with the same value in a specified column can be connected together by lines in the Scatter Plot. This helps identify groups of points and also visualize profiles using the scatter plot. The column specified must be a categorical column. This column will be used to group the points together. The order in which these will be connected by lines is given by another column, namely the 'Order By' column. This 'Order By' column can be categorical or continuous. See Figure 5.12

Labels: You can label each point in the plot by its value in a particular column; this column can be chosen in the Label Column drop-down list. Alternatively, you can choose to label only the selected points.

Fold Change Lines: This option allows the user to draw fold change lines on the scatter plot based on the following equations:

- $y = x + \log(\text{FC})$, $y = x$, $y = x - \log(\text{FC})$ ———(1)

- $y = (\text{FC}) x$, $y = x$, $y = (1/\text{FC}) x$ ———(2)

The lines that are drawn on the Scatter Plot depends on the following two parameters:

1. Datatype

- **Normalized Data:** If the scales chosen for the x-axis and y-axis is not same then no lines are drawn. If the scales chosen are same and are 'None' then lines given by (1)

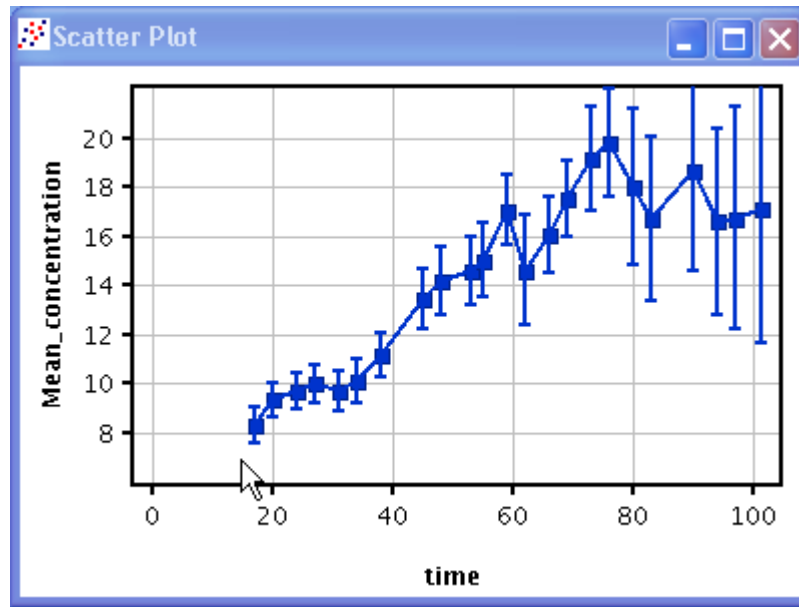


Figure 5.12: Viewing Profiles and Error Bars using Scatter Plot

above are drawn. If the scales chosen are 'Log' then no lines are drawn. If the scales chosen are 'Symmetric Log' then no lines are drawn.

- **Raw Data:** If the scales chosen for the x-axis and y-axis is not same then no lines are drawn. If the scales chosen are same and are 'None' then lines given by (2) above are drawn. If the scales chosen are 'Log' then lines given by (1) above are drawn. If the scales chosen are 'Symmetric Log' then no lines are drawn.

2. Axis Scale

The default fold change value given is 2.0. The user can change the default value either by moving the slider or entering the value in the appropriate box. When the default value is changed, the view gets dynamically altered to reflect the new fold change lines. See figure 5.13.

Rendering: The Scatter plot allows all aspects of the view to be customized. Fonts, colors, offsets, etcetera can all be configured.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the *Properties* dialog. To change a *Font*, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a *Color Chooser*. Select the desired color and click *OK*. This will change the corresponding color in the View.

Offsets: The bottom offset, top offset, left offset, and right offset of the plot can be modified and configured. These offsets may need to be changed if the axis labels or axis titles are not

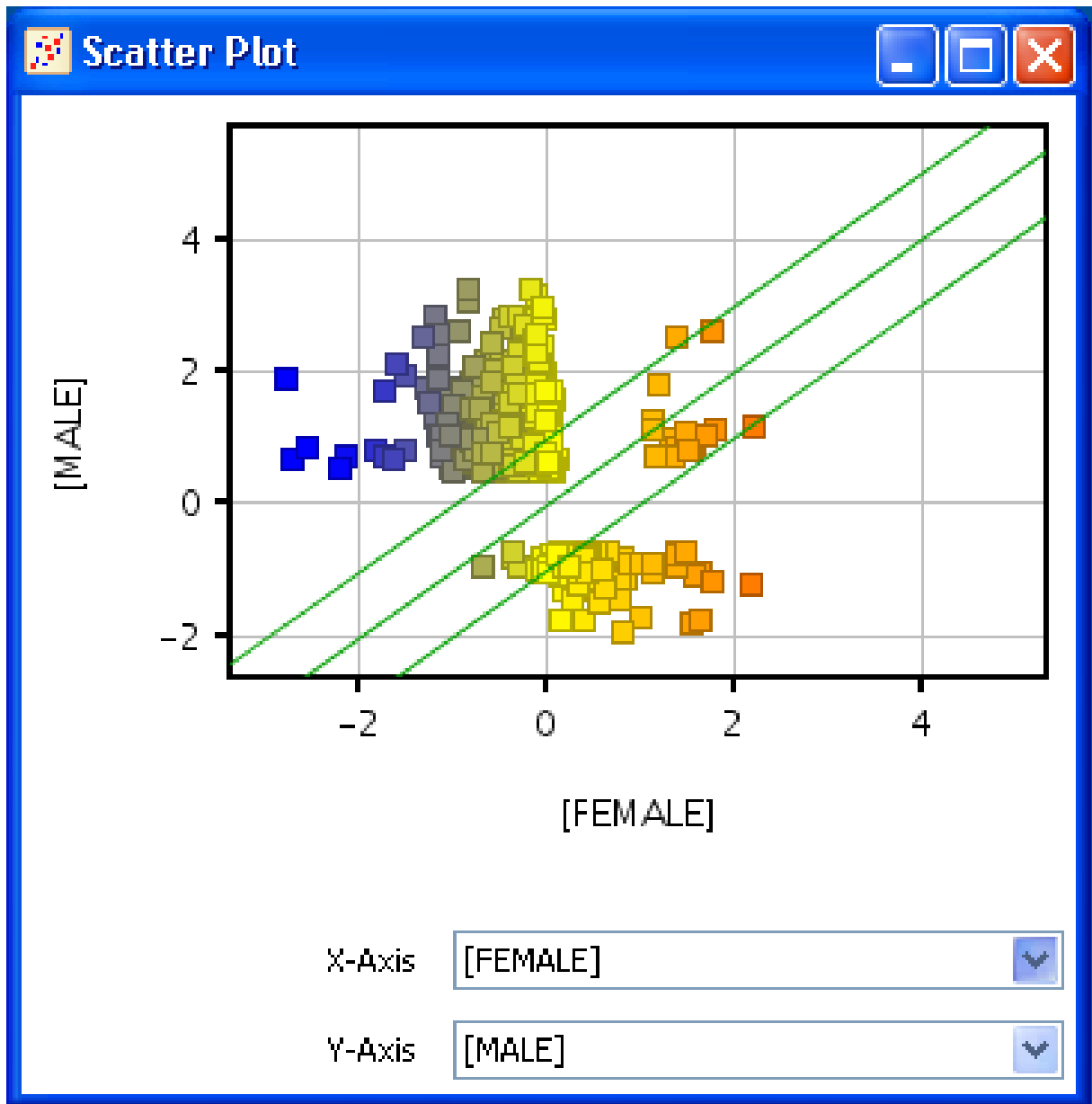


Figure 5.13: Scatter plot with Fold Change lines

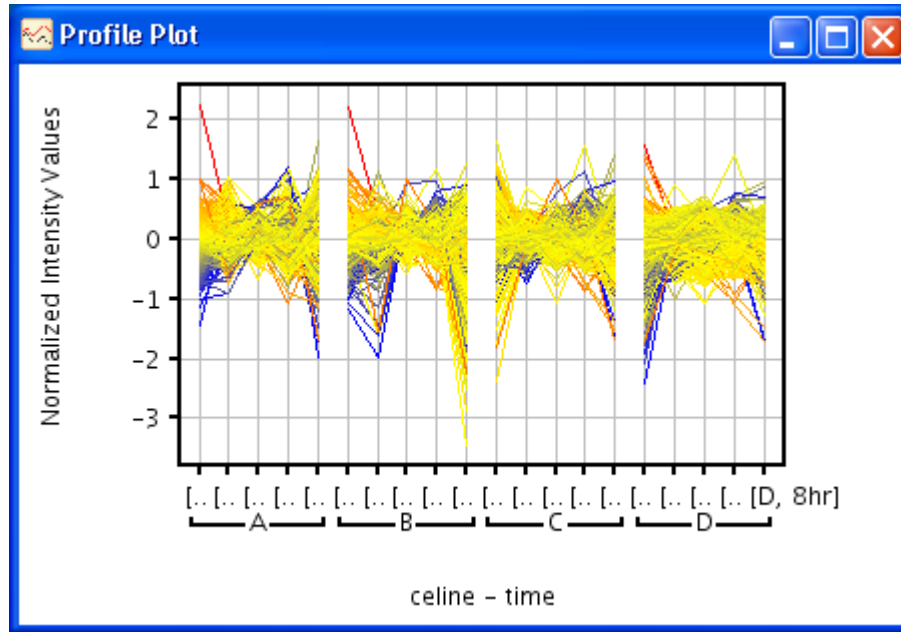


Figure 5.14: Profile Plot


completely visible in the plot, or if only the graph portion of the plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Miscellaneous: The quality of the plot can be enhanced by anti aliasing all the points in the plot. this is done to ensure better print quality. To enhance the plot quality, click on the High Quality Plot option.

Column Chooser: The column chooser can be disable and removed from the scatter plot if required. The plot area will be increased and the column chooser will not be available on the scatter plot. To remove the column chooser from the plot, uncheck the Show Column Chooser option.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.5 The Profile Plot View

The Profile Plot is launched from the view menu on the main menu bar. Alternately, Left-Click on the tool bar 'Profile Plot'  icon will bring up the profile plot. The profile plot (referred to as 'Graph View' in earlier versions of **GeneSpring GX**) is one of the important visualizations of normalized expression value data against the chosen interpretation. In fact, the default view of visualizing interpretations is the profile plot launched by clicking on the interpretation in the experiment and making it the active interpretation. See Figure 5.14

When the profile plot is launched from the view menu, it is launched with the active interpretation and the active entity list in the experiment. The profile plot shows the conditions in the active interpretation along the x-axis and the normalized expression values in the y-axis. Each entity in the active entity list is shown as a profile in the plot. Depending upon the interpretation, whether averaged or unaveraged, the profile of the entity in each group is split and displayed along the conditions in the interpretation.

Profile Plot for All Samples: If the active interpretation is the default *All Samples* interpretation, then each sample is shown in the x-axis and the normalized expression values for each entity in the active entity list is connected across all the samples.

Profile Plot of Unaveraged Interpretation: If the active interpretation is *unaveraged* over the replicates, then the samples in each condition are grouped together along the x-axis, and the profile plot of the entities in the active interpretation is continuous within the samples in a condition and split across the conditions.

Profile Plot of Averaged Interpretation: If the active interpretation is averaged, over the replicates, then the conditions in the interpretation are plotted on the x-axis. The profile plot of the entities in the active entity list is displayed continuously with the averaged condition. And if there are multiple parameters in the interpretation, the profile plot will be split by the outer most parameter. Thus if the first parameter is dosage and the second parameter is Gender (Male and Female), and these two parameters combine to make conditions, then the profile will be continuous with dosage and split along Gender.

Clicking on another entity list in the experiment will make that entity list active and the profile plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the profile plot.

The Profile Plot supports both the Selection Mode and the Zoom Modes The profile plot is launched with the selection mode as default and colored by the values in the first condition. The interpretation of the profile plot and the color band are displayed in the legend window.

5.5.1 Profile Plot Operations

The Profile Plot operations are accessed by right-clicking on the canvas of the Profile Plot. Operations that are common to all views are detailed in the section [Common Operations on Plot Views](#). Profile Plot specific operations and properties are discussed below.

Selection Mode: The Profile Plot is launched, by default, in the selection mode. While in the selection mode, left-clicking and dragging the mouse over the Profile Plot will draw a selection box and all profiles that intersect the selection box are selected. To select additional profiles, Ctrl-Left-Click and drag the mouse over desired region. Individual profiles can be selected by clicking on the profile of interest.

Zoom Mode: While in the zoom mode, left-clicking and dragging the mouse over the selected region draws a zoom box and will zoom into the region. Reset Zoom will revert back to the default, showing the plot for all the entities in the active entity list.

5.5.2 Profile Plot Properties

The following properties are configurable in the Profile Plot. See [Figure 5.15](#)

Axis: The grids, axes labels, and the axis ticks of the plots can be configured and modified. To modify these, Right-Click on the view, and open the Properties dialog. Click on the Axis tab. This will open the axis dialog.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For continuous data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Visualization: The Profile Plot displays the mean profile over all rows by default. This can be hidden by unchecking the Display Mean Profile check box.

The colors of the Profile Plot can be changed from the properties dialog. The colors of the profile plot can be changed from this dialog. You can choose a fixed color or use one of the data columns to color the profile plot by choosing a column from the drop-down list. The colors range of the profile plot and the middle color can be customized by clicking on the *Customize* button and choosing the minimum color, the middle color and the maximum color. By default, the minimum color is set to the median value of the data column.

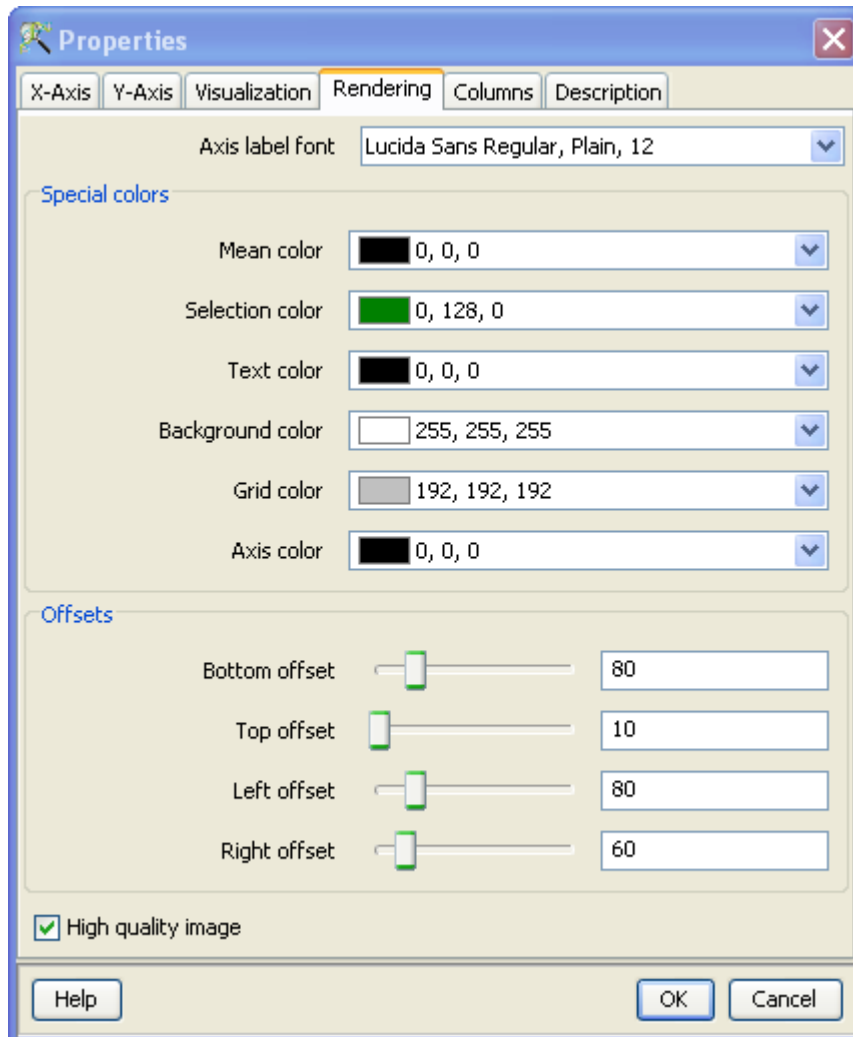


Figure 5.15: Profile Plot Properties

Rendering: The rendering of the fonts, colors and offsets on the Profile Plot can be customized and configured.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the *Properties* dialog. To change a *Font*, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a *Color Chooser*. Select the desired color and click *OK*. This will change the corresponding color in the View.

Offsets: The bottom offset, top offset, left offset, and right offset of the plot can be modified and

configured. These offsets may be need to be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of the plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Quality Image: The Profile Plot image quality can be increased by checking the High-Quality anti-aliasing option. This is slow however and should be used only while printing or exporting the Profile Plot.

Column: The Profile Plot is launched with a default set of columns. The set of visible columns can be changed from the Columns tab. The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the Selected items to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the Selected items list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.


To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.

The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.6 The Heatmap View

The heatmap is launched from View Menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar 'Heatmap'  icon will bring up the heat map view. The Heat Map displays the normalized signal values of the conditions in the active interpretation for all the entities in the active entity list. The legend window displays the interpretation on which the heat map was launched.

Clicking on another entity list in the experiment will make that entity list active and the heatmap will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the heat map.

The expression value of each gene is mapped to a color-intensity value. The mapping of expression values to intensities is depicted by a color-bar created by the range of values in the conditions of the interpretation. This provides a birds-eye view of the values in the dataset. The tool tip on a cell in the heat map shows the normalized expression value of the entity.

The heat map allows selecting the entities (rows) and selecting the conditions (columns) and these are lassoed in all the views. To select contiguous cells in the heat map, click and drag to draw a rectangular box on the canvas of the heat map. The corresponding entities and conditions will be selected and lassoed across all views. See Figure [5.16](#)

5.6.1 Heatmap Operations

Operations on heat map are accessible on the tool bar and Right-Click menu on the canvas of the heat map. Operations that are common to all views are detailed in the section [Common Operations on Table Views](#) above. In addition, some of the heat specific operations and the HeatMap properties are explained below: See Figure [5.17](#)

Export As Image: This will pop-up a dialog to export the view as an image. This functionality allows the user to export very high quality image. You can specify any size of the image, as well as the

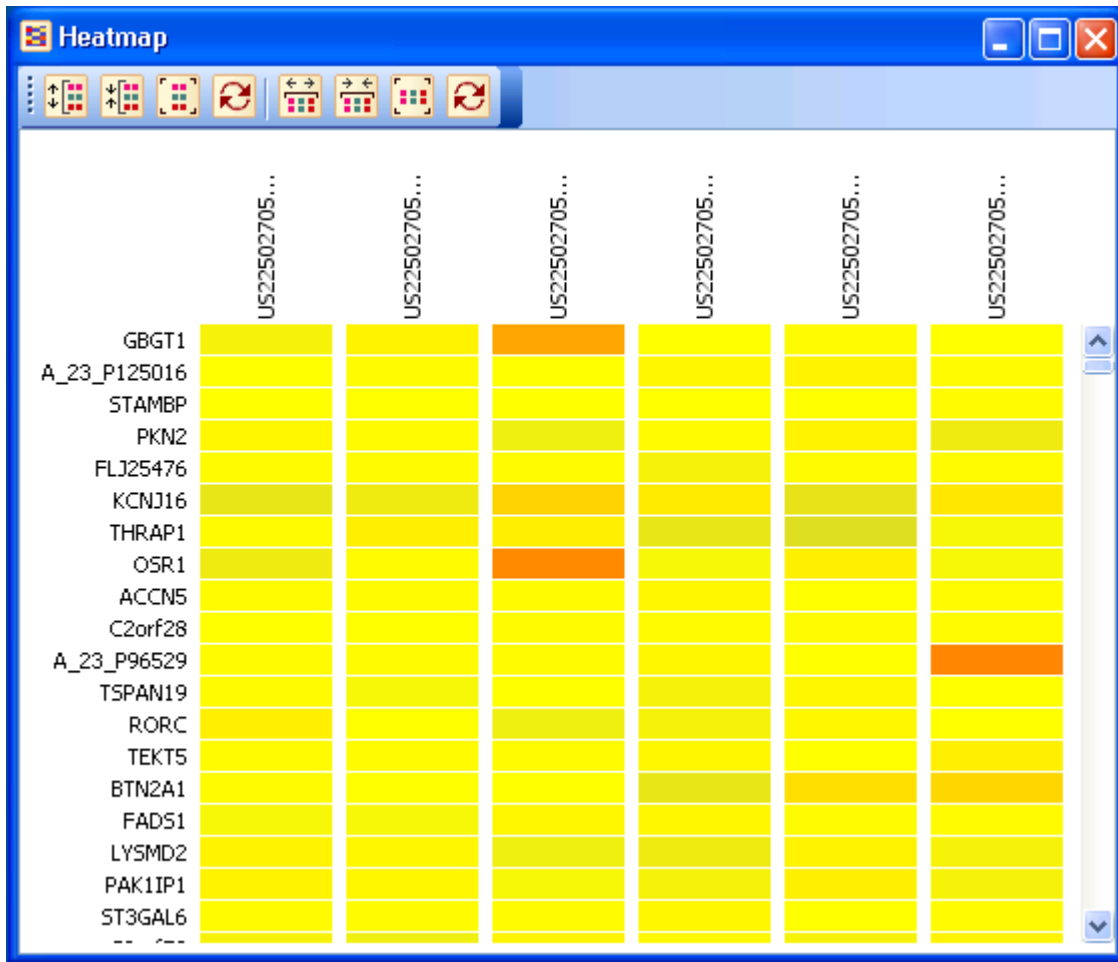


Figure 5.16: Heat Map

resolution of the image by specifying the required dots per inch (dpi) for the image. Images can be exported in various formats. Currently supported formats include png, jpg, jpeg, bmp or tiff. Finally, images of very large size and resolution can be printed in the tiff format. Very large images will be broken down into tiles and recombined after all the images pieces are written out. This ensures that memory is but built up in writing large images. If the pieces cannot be recombined, the individual pieces are written out and reported to the user. However, tiff files of any size can be recombined and written out with compression. The default dots per inch is set to 300 dpi and the default size if individual pieces for large images is set to 4 MB. These default parameters can be changed in the tools → Options dialog under the **Export as Image**

The user can export only the visible region or the whole image. Images of any size can be exported with high quality. If the whole image is chosen for export, however large, the image will be broken up into parts and exported. This ensures that the memory does not bloat up and that the whole high quality image will be exported. After the image is split and written out, the tool will attempt to combine all these images into a large image. In the case of png, jpg, jpeg and bmp often this will not be possible because of the size of the image and memory limitations. In such cases, the individual images will be written separately and reported. However, if a tiff image format is chosen, it will be exported as a single image however large. The final tiff image will be compressed and saved.

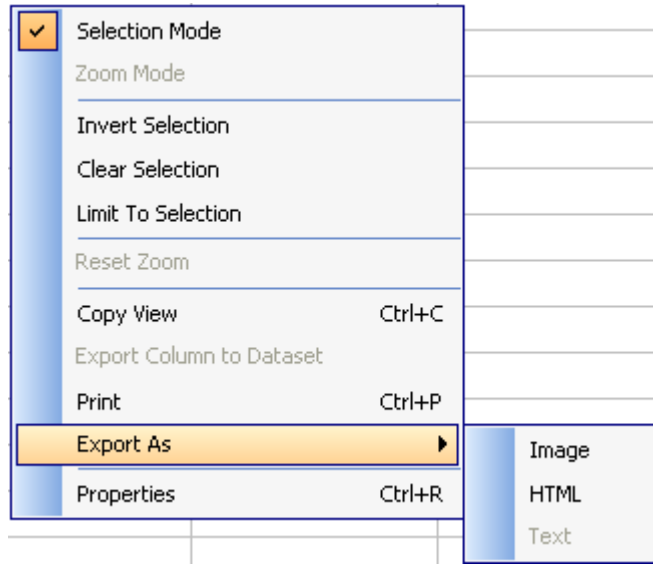


Figure 5.17: Export submenus

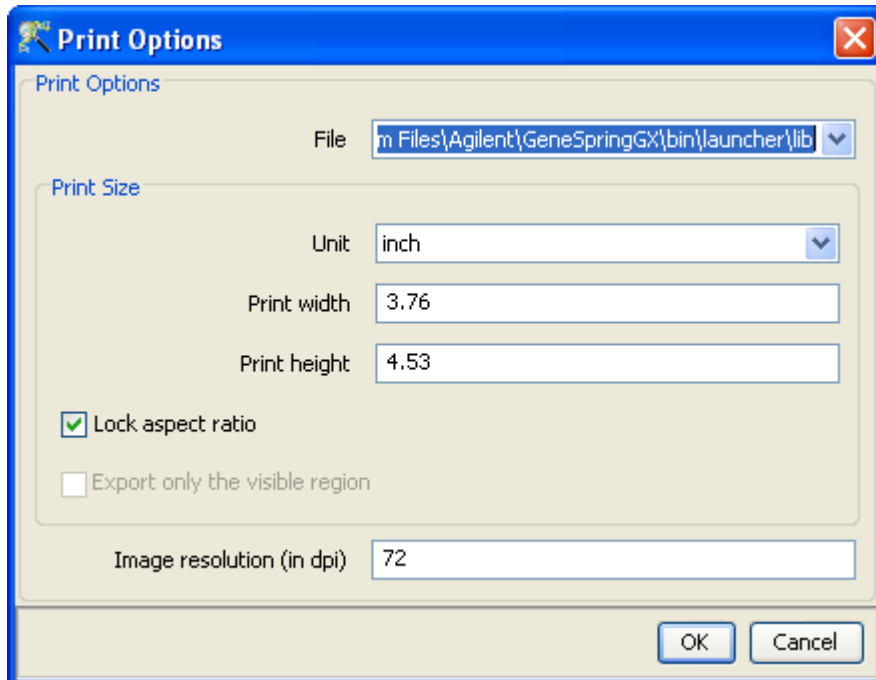


Figure 5.18: Export Image Dialog

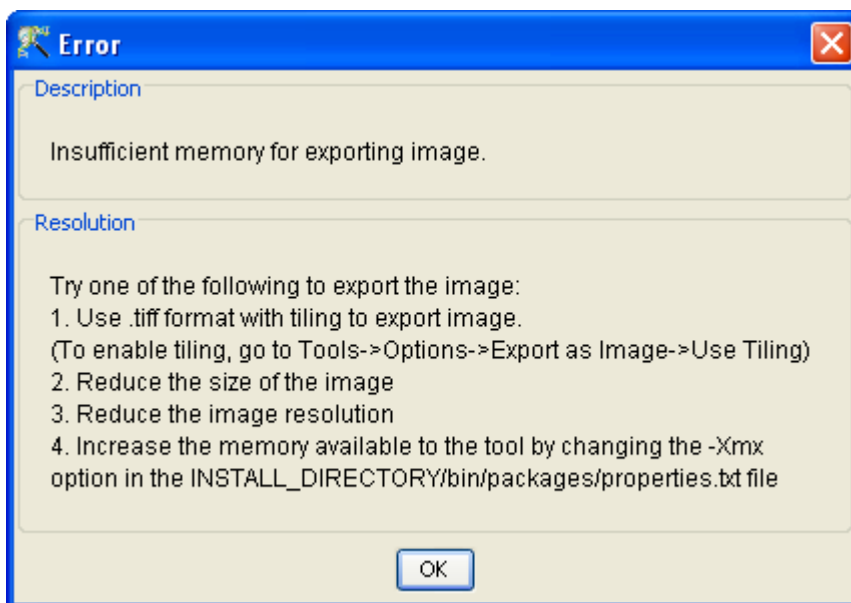


Figure 5.19: Error Dialog on Image Export

Note: This functionality allows the user to create images of any size and with any resolution. This produces high-quality images and can be used for publications and posters. If you want to print very large images or images of very high-quality the size of the image will become very large and will require huge resources. If enough resources are not available, an error and resolution dialog will pop up, saying the image is too large to be printed and suggesting you to try the tiff option, reduce the size of image or resolution of image, or to increase the memory available to the tool by changing the -Xmx option in `INSTALL_DIR/bin/packages/properties.txt` file. On **Mac OS X** the Java heap size parameters are set in in the file `Info.plist` located in `INSTALL_DIR/GeneSpringGX.app/Contents/Info.plist`. Change the Xmx parameter appropriately. Note that in the Java heap size limit on Mac OS X is about 2048M.

Note: You can export the whole heatmap as a single image with any size and desired resolution. To export the whole image, choose this option in the dialog. The whole image of any size can be exported as a compressed tiff file. This image can be opened on any machine with enough resources for handling large image files.

Export as HTML: This will export the view as an html file. Specify the file name and the the view will be exported as an HTML file that can be viewed in a browser and deployed on the web. If the whole image export is chosen, multiple images will be exported and can be opened in a browser.



Figure 5.20: heatmap Toolbar

5.6.2 Heatmap Toolbar

The icons on the heatmap and their operations are listed below: See [Figure 5.20](#)



Expand rows: Click to increase the row dimensions of the heatmap. This increases the height of every row in the heatmap. Row labels appear once the inter-row separation is large enough to accommodate label strings.



Contract rows: Click to reduce row dimensions of the heatmap so that a larger portion of the heatmap is visible on the screen.



Collapse Rows: Click to scale the rows of the heatmap to fit entirely in the window. A large image, which needs to be scrolled to view completely, fails to effectively convey the entire picture. Fitting it to the screen gives an overview of the whole dataset.



Reset rows: Click to scale the heatmap back to default resolution showing all the row labels.

Note: Row labels are not visible when the spacing becomes too small to display labels. Zooming in or Resetting will restore these.



Expand columns: Click to scale up the heatmap along the columns.



Contract columns: Click to reduce the scale of the heatmap along columns. The cell width is reduced and more of the heatmap is visible on the screen.

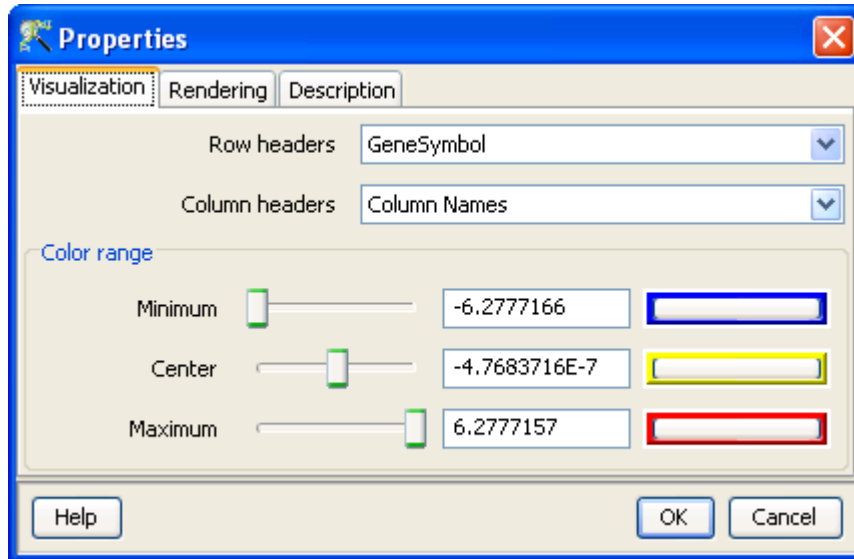


Figure 5.21: heatmap Properties



Collapse columns: Click to scale the columns of the heatmap to fit entirely in the window. This is useful in obtaining an overview of the whole dataset. A large image, which needs to be scrolled to view completely, fails to effectively convey the entire picture. Fitting it to the screen gives a quick overview.



Reset columns: Click to scale the heatmap back to default resolution. Note: Column Headers are not visible when the spacing becomes too small to display labels. Zooming or Resetting will restore these.

5.6.3 heatmap Properties

The heatmap views supports the following configurable properties. See Figure 5.21

Visualization: Row headers: Any annotation column can be used to label the rows of the heatmap from the **Row headers** drop down list.

Column headers: The column headers on the heatmap is labeled with the names of the interpretation on which the heatmap is launched. If all samples are used, or an unaveraged interpretation is used, the column headers show the column names. If column headers are not required, they can set to **None** from the drop-down list.

Color range: The Color and Saturation Threshold of the heatmap can be changed from the Properties Dialog. The saturation threshold can be set by the Minimum, Center and Maximum sliders

or by typing a numeric value into the text box and hitting Enter. The colors of Minimum, Center and Maximum can be set from the corresponding color chooser dialog. All values above the Maximum and values below the Minimum are thresholded to Maximum and Minimum colors respectively. The chosen colors are graded and assigned to cells based on the numeric value of the cell. Values between maximum and center are assigned a graded color in between the extreme maximum and center colors, and likewise for values between minimum and center.

Rendering: The rendering of the heatmap can be customized and configured from the rendering tab of the heatmap properties dialog.

The location of the row and column headers can be set from the drop-down list.

The row and column labels are shown along with the heatmap. These widths allotted for these labels can be configured.


The default vertical and horizontal spacing of the cells of the heat map can be changed.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.6.4 Heatmap for viewing Copy Number Analysis Results

Heatmap view is supported for visualizing the results of Copy Number Analysis in **GeneSpring GX**. Copy Number and LOH values can be visualized in the heatmap for a chosen entity list and interpretation. See Section [Heatmap View for Copy Number](#).

5.7 The Histogram View

The Histogram is launched from View menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar 'Histogram'  icon will bring up the histogram. This toolbar provides the option to view either a single histogram or multiple histograms in one view. Multiple histograms will be launched with different samples in the experiment and there is an option in 'Properties → Rendering' to set the number of samples to be shown in the view.

The view shows a histogram of one condition in the active interpretation as a bar chart of the frequency or number of entities in each interval of the condition. This is done by binning the normalized signal value of the condition into equal interval bins and plotting the number of entities in each bin. If the default *All Samples* interpretation is chosen, the histogram will correspond to the normalized signal values of the first sample. If an averaged interpretation is active interpretation, then the histogram will correspond to

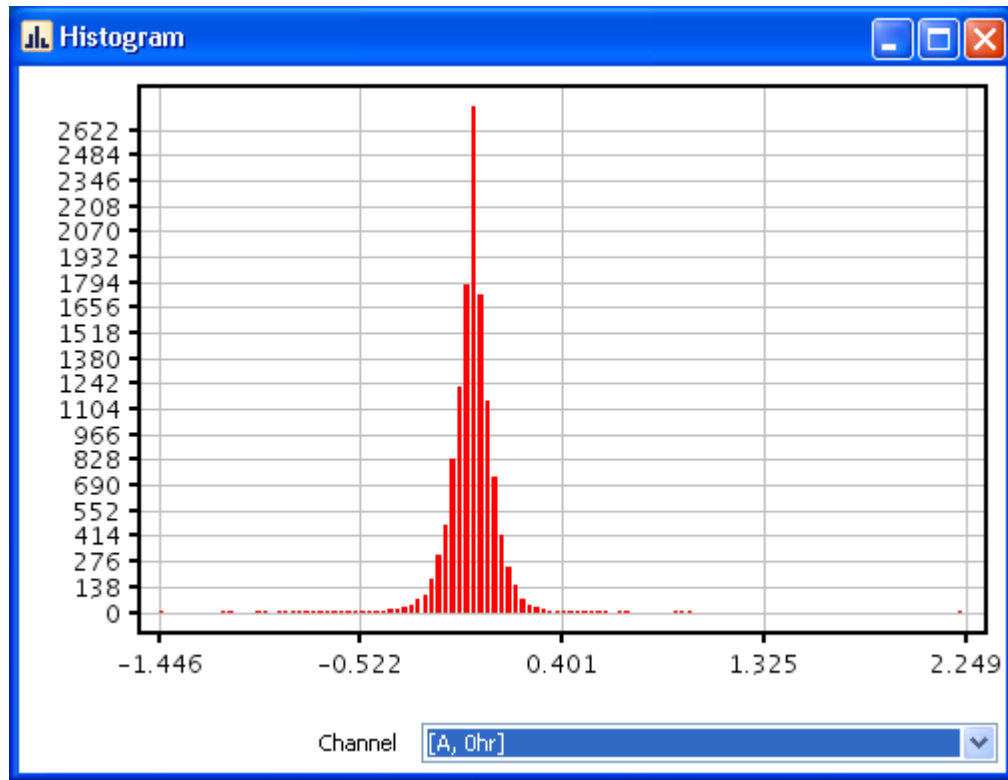


Figure 5.22: Histogram

the averaged normalized signal values of the samples in the first condition. You can change the condition on which the histogram is drawn from the drop-down list on the view. The legend window displays the interpretation on which the histogram was launched. See [Figure 5.22](#)

Clicking on another entity list in the experiment will make that entity list active and the histogram will dynamically display the frequency of this entity list on the condition. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display the frequency of those entities in the histogram.

The frequency in each bin of the histogram is dependent upon the lower and upper limits of binning, and the size of each bin. These can be configured and changed from the **Properties** dialog.

When 'multiple histogram' is launched, the view shows the histogram of multiple samples (as set in the Number of samples options in 'Rendering' section of **Properties** dialog. In this case, the setting in **Properties** dialog apply to all the samples and hence all the histograms.

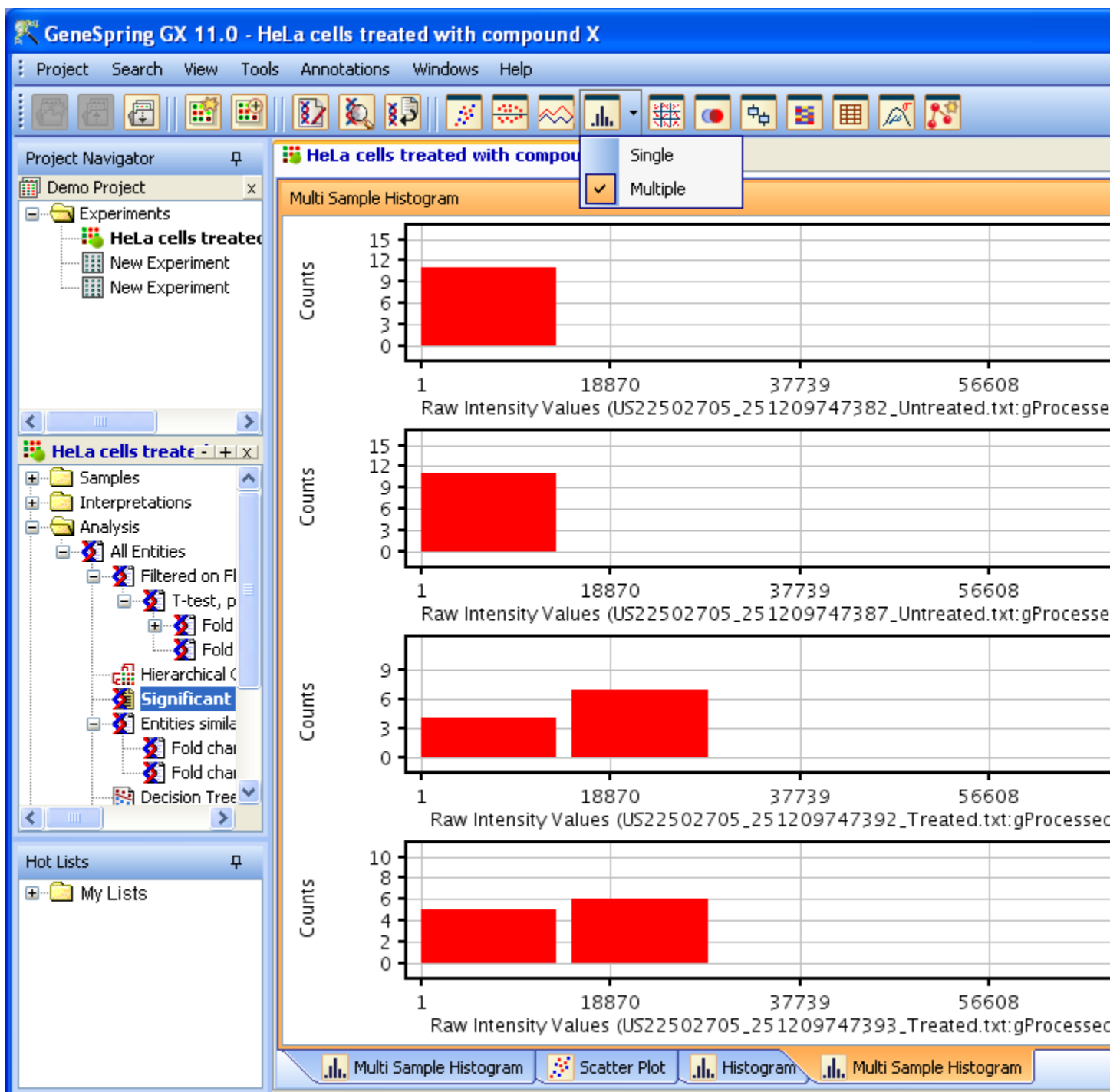


Figure 5.23: Histogram Viewing Options

5.7.1 Histogram Operations

The Histogram operations are accessed by Right-Click on the canvas of the Histogram Operations that are common to all views are detailed in the section [Common Operations on Plot Views](#). Histogram-specific operations and properties are discussed below.

Selection Mode: The Histogram supports only the Selection mode. Left-Click and dragging the mouse over the Histogram draws a selection box and all bars that intersect the selection box are selected and lassoed. Clicking on a bar also selects the elements in that bar. To select additional elements, Ctrl-Left-Click and drag the mouse over the desired region.

5.7.2 Histogram Properties

The Histogram can be viewed with different channels, user-defined binning, different colors, and titles and descriptions from the Histogram Properties Dialog. See Figure [5.24](#)

The Histogram Properties Dialog is accessible by right-clicking on the histogram and choosing **Properties** from the menu. The histogram view can be customized and configured from the histogram properties.

Axis: The histogram channel can be changed from the Properties menu. Any column in the dataset can be selected here.

The grids, axes labels, and the axis ticks of the plots can be configured and modified. To modify these, Right-Click on the view, and open the Properties dialog. Click on the Axis tab. This will open the axis dialog.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For continuous data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Visualization: Color By: You can specify a Color By column for the histogram. The Color By should be a categorical column in the active dataset. This will color each bar of the histogram with different color bars for the frequency of each category in the particular bin.

Explicit Binning: The Histogram is launched with a default set of equal interval bins for the chosen column. This default is computed by dividing the interquartile range of the column values into

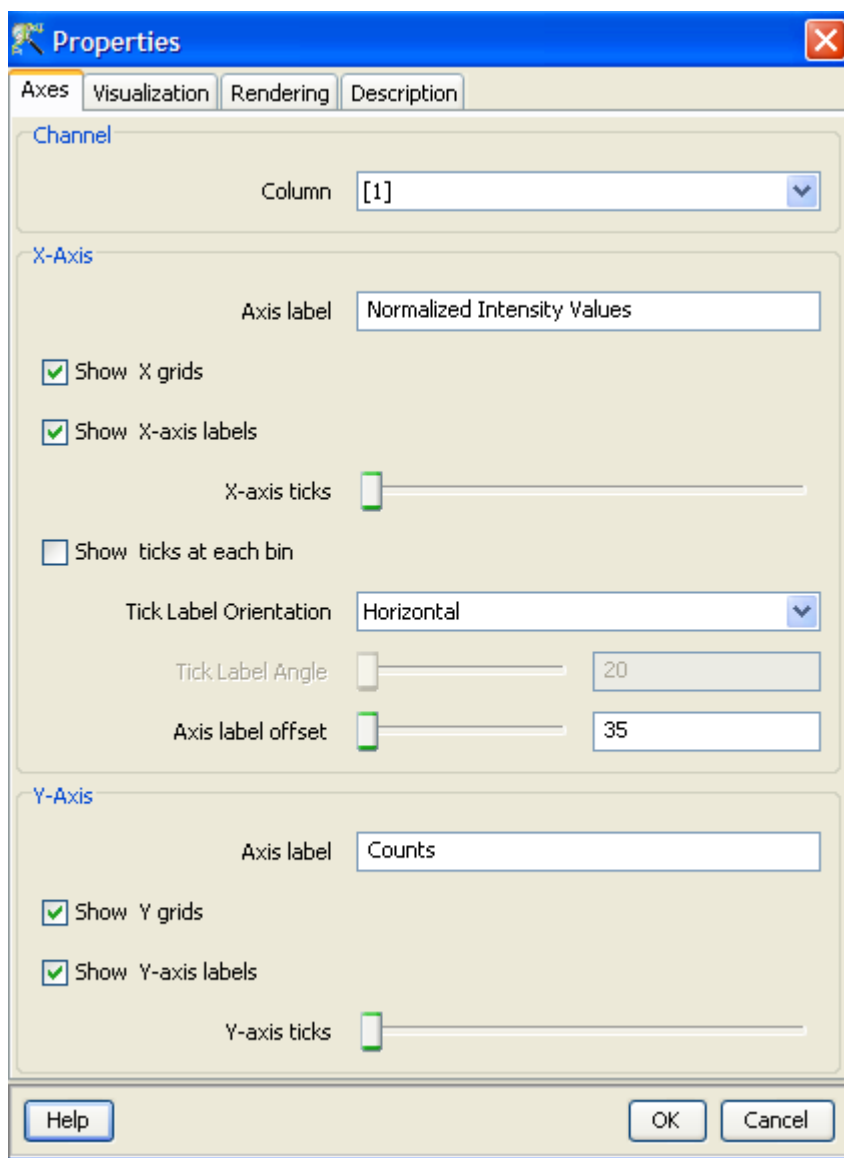


Figure 5.24: Histogram Properties

three bins and expanding these equal interval bins for the whole range of data in the chosen column. The Histogram view is dependent upon binning and the default number of bins may not be appropriate for the data. The data can be explicitly re-binned by checking the Use Explicit Binning check box and specifying the minimum value, the maximum value and the number of bins using the sliders. The maximum - minimum values and the number of bins can also be specified in the text box next to the sliders. Please note that if you type values into the text box, you will have to hit Enter for the values to be accepted.

Bar Width: the bar width of the histogram can be increased or decreased by moving the slider. The default is set to 0.9 times the area allocated to each histogram bar. This can be reduced if desired.

Channel chooser: The Channel Chooser on the histogram view can be disabled by unchecking the check box. This will afford a larger area to view the histogram.

Rendering: This tab provides the interface to customize and configure the fonts, the colors and the offsets of the plot.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the *Properties* dialog. To change a *Font*, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a *Color Chooser*. Select the desired color and click *OK*. This will change the corresponding color in the View.

Offsets: The bottom offset, top offset, left offset, and right offset of the plot can be modified and configured. These offsets may be need to be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of the plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

There is also a provision to set the number of samples that can be seen in the view, if the histogram tool bar was launched with the 'Multiple' option.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

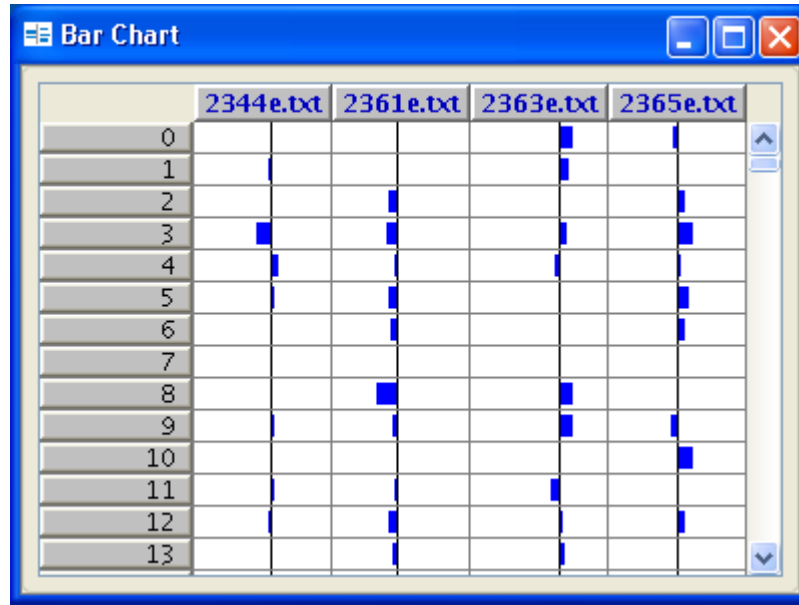


Figure 5.25: Bar Chart

5.8 The Bar Chart

The Bar Chart is launched from a script with the default interpretation. `script.view.BarChart().show()` By default, the Bar Chart is launched with all continuous columns in the active dataset. The Bar Chart provides a view of the range and distribution of values in the selected column. The Bar Chart is a tabular view and thus all operations that are possible on a table are possible here. The Bar Chart can be customized and configured from the **Properties** dialog accessed from the Right-Click menu on the canvas of the Chart. See Figure 5.25

Note that the Bar Chart will show only the continuous columns in the current dataset.


5.8.1 Bar Chart Operations

The Operations on the Bar Chart is accessible from the menu on Right-Click on the canvas of the Bar Chart. Operations that are common to all views are detailed in the section [Common Operations on Table Views](#) above. In addition, some of operations and the bar chart properties are explained below:

Sort: The Bar Chart can be used to view the sorted order of data with respect to a chosen column as bars. Sort is performed by clicking on the column header. Mouse clicks on the column header of the bar chart will cycle through an ascending values sort, a descending values sort and a reset sort. The column header of the sorted column will also be marked with the appropriate icon.

Thus to sort a column in the ascending order, click on the column header. This will sort all rows

of the bar chart based on the values in the chosen column. Also an icon on the column header will denote that this is the sorted column. To sort in the descending order, click again on the same column header. This will sort all the rows of the bar chart based on the decreasing values in this column. To reset the sort, click again on the same column. This will reset the sort and the sort icon will disappear from the column header.

Selection: The bar chart can be used to select rows, columns, or any contiguous part of the dataset. The selected elements can be used to create a subset dataset by left-clicking on Create dataset from Selection  icon.

Row Selection: Rows are selected by left-clicking on the row headers and dragging along the rows. Ctrl-Left-Click selects subsequent items and Shift-Left-Click selects a consecutive set of items. The selected rows will be shown in the lasso window and will be highlighted in all other views.

Column Selection: Columns can be selected by left-clicking in the column of interest. Ctrl-Left-Click selects subsequent columns and Shift-Left-Click selects consecutive set of columns. The current column selection on the bar chart usually determines the default set of selected columns used when launching any new view, executing commands or running algorithm. The selected columns will be lassoed in all relevant views and will be show selected in the lasso view.

5.8.2 Bar Chart Properties

The Bar Chart Properties Dialog is accessible by Right-Click on the bar chart and choosing **Properties** from the menu. The bar chart view can be customized and configured from the bar chart properties.

Rendering: The rendering tab of the bar chart dialog allows you to configure and customize the fonts and colors that appear in the bar chart view.

Special Colors: All the colors in the Table can be modified and configured. You can change the Selection color, the Double Selection color, Missing Value cell color and the Background color in the table view. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the properties dialog. To change a color, click on the appropriate color bar. This will pop-up a Color Chooser. Select the desired color and click OK. This will change the corresponding color in the Table.

Fonts: Fonts that occur in the table can be formatted and configured. You can set the fonts for Cell text, row Header and Column Header. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a Font, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Visualization: The display precision of decimal values in columns, the row height, the missing value text, and the facility to enable and disable sort are configured and customized by options in this tab.

The visualization of the display precision of the numeric data in the table, the table cell size and the text for missing value can be configured. To change these, Right-Click on the table view and open the Properties dialog. Click on the visualization tab. This will open the Visualization panel.

To change the numeric precision. Click on the drop-down box and choose the desired precision. For decimal data columns, you can choose between full precision and one to four decimal places, or representation in scientific notation. By default, full precision is displayed.

You can set the row height of the table, by entering a integer value in the text box and pressing Enter. This will change the row height in the table. By default the row height is set to 16.

You can enter any a text to show missing values. All missing values in the table will be represented by the entered value and missing values can be easily identified. By default all the missing value text is set to an empty string.

You can also enable and disable sorting on any column of the table by checking or unchecking the check box provided. By default, sort is enabled in the table. To sort the table on any column, click on the column header. This will sort the all rows of the table based on the values in the sort column. This will also mark the sorted column with an icon to denote the sorted column. The first click on the column header will sort the column in the ascending order, the second click on the column header will sort the column in the descending order, and clicking the sorted column the third time will reset the sort.

Columns: The order of the columns in the bar chart can be changed by changing the order in the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the *Selected items* to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the *Selected items* list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.

To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.

The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

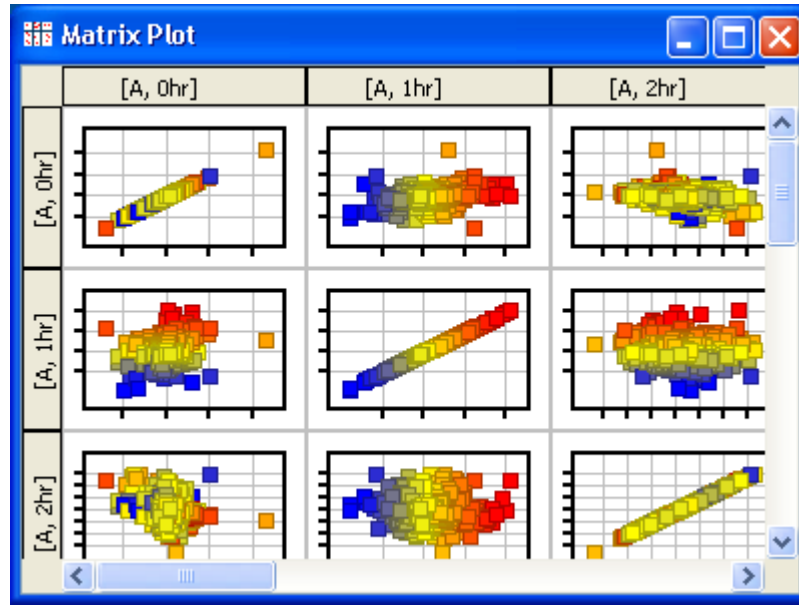



Figure 5.26: Matrix Plot

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.9 The Matrix Plot View

The Matrix Plot is launched from the View menu on the main menu bar with the active interpretation and the active entity list. Alternately, Left-Click on the tool bar 'Matrix plot'  icon will bring up the Matrix plot. The Matrix Plot shows a matrix of pairwise 2D scatter plots for conditions in the active interpretation. The X-Axis and Y-Axis of each scatter plot corresponding to the conditions in the active interpretation are shown in the corresponding row and column of the matrix plot. See Figure 5.26

If the active interpretation is the default *All Samples* interpretation, the matrix plot shows the normalized expression values of each sample against the other. If an averaged interpretation is the active interpretation, then the matrix plot will show the averaged normalized signal values of the samples in each condition against the other. The points in the matrix plot correspond to the entities in the active entity list. The legend window displays the interpretation on which the matrix plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the matrix plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the matrix plot.

The main purpose of the matrix plot is to get an overview of the correlation between conditions in the dataset, and detect conditions that separate the data into different groups.

By default, a maximum of 10 conditions can be shown in the matrix plot. If more than 10 conditions are present in the active interpretation, only ten conditions are projected into the matrix plot and other columns are ignored with a warning message. The matrix plot is interactive and can be lassoed. Elements of the matrix plot can be configured and altered from the properties menu described below.

5.9.1 Matrix Plot Operations

The Matrix Plot operations are accessed from the main menu bar when the plot is the active windows. These operations are also available by right-clicking on the canvas of the Matrix Plot. Operations that are common to all views are detailed in the section [Common Operations on Plot Views](#). Matrix Plot specific operations and properties are discussed below.

Selection Mode: The Matrix Plot supports only the Selection mode. Left-Click and dragging the mouse over the Matrix Plot draws a selection box and all points that intersect the selection box are selected and lassoed. To select additional elements, Ctrl-Left-Click and drag the mouse over the desired region. Ctrl-Left-Click toggles selection. This selected points will be unselected and unselected points will be added to the selection and lassoed.

5.9.2 Matrix Plot Properties

The matrix plot can be customized and configured from the properties dialog accessible from the Right-Click menu on the canvas of the Matrix plot. The important properties of the scatter plot are all available for the Matrix plot. These are available in the Axis tab, the Visualization tab, the Rendering tab, the Columns tab and the description tab of the properties dialog and are detailed below. See Figure [5.27](#)

Axis: The Axes on the Matrix Plot can be toggled to show or hide the grids, or show and hide the axis

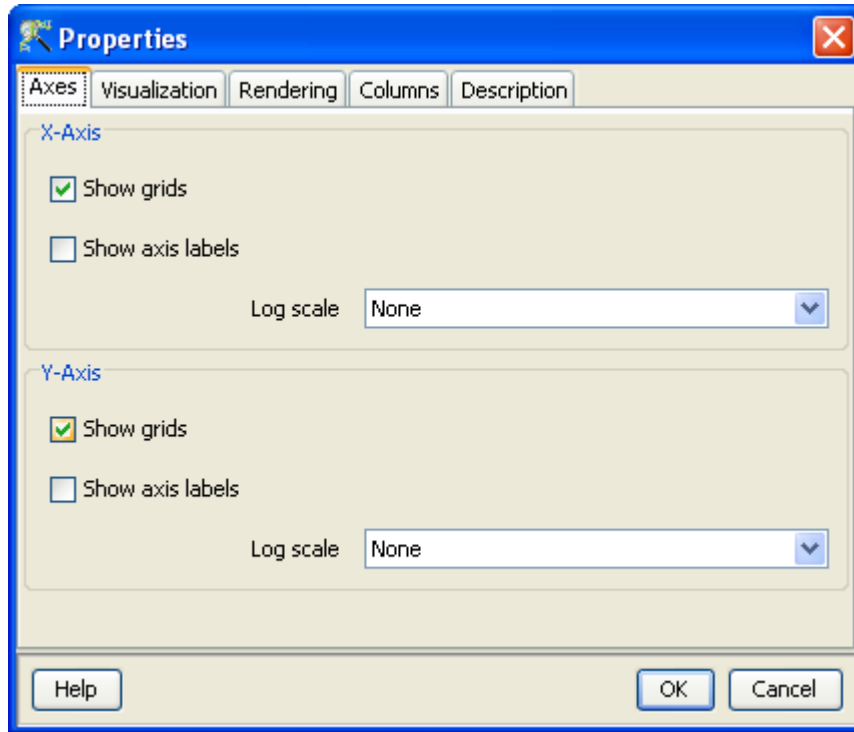


Figure 5.27: Matrix Plot Properties

labels.

Visualization: The scatter plots can be configured to Color By any column of the active dataset, Shape By any categorical column of the dataset, and Size by any column of the dataset.

Rendering: The fonts on the Matrix Plot, the colors that occur on the Matrix Plot, the Offsets, the Page size of the view and the quality of the Matrix Plot can be altered from the Rendering tab of the Properties dialog.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the *Properties* dialog. To change a *Font*, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a *Color Chooser*. Select the desired color and click *OK*. This will change the corresponding color in the View.

Offsets: The bottom offset, top offset, left offset, and right offset of the plot can be modified and configured. These offsets may be need to be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of the plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab.

To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Page: The visualization page of the Matrix Plot can be configured to view a specific number of scatter plots in the Matrix Plot. If there are more scatter plots in the Matrix plot than in the page, scroll bars appear and you can scroll to the other plot of the Matrix Plot.

Plot Quality: The quality of the plot can be enhanced to be anti-aliased. This will produce better points and will produce better prints of the Matrix Plot.

Columns: The Columns for the Matrix Plot can be chosen from the Columns tab of the Properties dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the Selected items to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the Selected items list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.


To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.

The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.10 Summary Statistics View

The Summary Statistics View is launched from view menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar 'Summary Statistics'  icon will display the summary statistics. This view shows the summary statistics of the conditions in the active interpretation with respect to the active entity list. Thus, each column of the summary statistics shows the mean, standard deviation, median, percentiles and outliers of the conditions in the active interpretation with active entity list. In **GeneSpring GX**, points that lie outside the quartiles i.e., 25th percentile value- $1.5 * (\text{interquartile range})$ and 75th percentile value + $1.5 * (\text{interquartile range})$ are considered outliers. The interquartile range is between 75th percentile and 25th percentile.

Let's say you have 100 values in your dataset. If you sort them in ascending order, the 25th value is 4, and the 75th value is 7. Therefore, the interquartile range is $7 - 4 = 3$. $1.5 * \text{interquartile range} = 1.5 * 3 = 4.5$. Therefore, all values in the dataset which are less than or equal to $4 - 4.5 = -0.5$ and all values which are more than or equal to $7 + 4.5 = 11.5$ are considered as outliers.

If the active interpretation is the default *All Samples* interpretation, the table shows the summary statistics of each sample with respect to the active entity list. If an averaged interpretation is the active interpretation, the table shows the summary statistics of the conditions in the averaged interpretation with respect to the active entity list. The legend window displays the interpretation on which the summary statistics was launched.

Clicking on another entity list in the experiment will make that entity list active and the summary statistics table will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the summary statistics table.

This Summary Statistics View is a tabular view and thus all operations that are possible on a table are possible here. The summary statistics table can be customized and configured from the **Properties** dialog accessed from the Right-Click menu on the canvas of the Chart. See Figure [5.28](#)

This view presents descriptive statistics information on the active interpretation, and is useful to compare the distributions of different conditions in the interpretation.

	[A, 0hr]	[A, 1hr]
No. of Observations	12488	12488
No. of Missing Values	0	0
Minimum	-1.4460182	-0.91799164
Maximum	2.2489069	1.0259538
Mean	-5.164427E-4	-0.011688839
Trimmed Mean	.8272578E-4	-0.013146568
Median	0.001390934	-0.008618355
Std. Deviation	0.1067329	0.12283408
Trimmed Std. Deviation	0.08228829	0.10526155
No. Of Outliers	672	512
Percentile 1.0	-0.2912606	-0.30164707
Percentile 5.0	-0.15345882	-0.20199674
Percentile 10.0	-0.10685689	-0.15382889

Figure 5.28: Summary Statistics View

5.10.1 Summary Statistics Operations

The Operations on the Summary Statistics View are accessible from the menu on Right-Click on the canvas of the Summary Statistics View. Operations that are common to all views are detailed in the section [Common Operations on Table Views](#) above. In addition, some of the Summary Statistics View specific operations and the bar chart properties are explained below:

Column Selection: The Summary Statistics View can be used to select conditions or columns. The selected columns are lassoed in all the appropriate views.

Columns can be selected by left-clicking in the column of interest. Ctrl-Left-Click selects subsequent columns and Shift-Left-Click consecutive set of columns. The current column selection on the bar chart usually determines the default set of selected columns used when launching any new view, executing commands or running algorithms. The selected columns will be lassoed in all relevant views and will be shown selected in the lasso view.

Export As Text: The *Export* → *Text* option saves the tabular output to a tab-delimited file that can be opened in **GeneSpring GX**.

5.10.2 Summary Statistics Properties

The Summary Statistics View Properties Dialog is accessible by right-clicking on the Summary Statistics View and choosing **Properties** from the menu. The Summary Statistics View can be customized and configured from the Summary Statistics View properties. See [Figure 5.29](#)

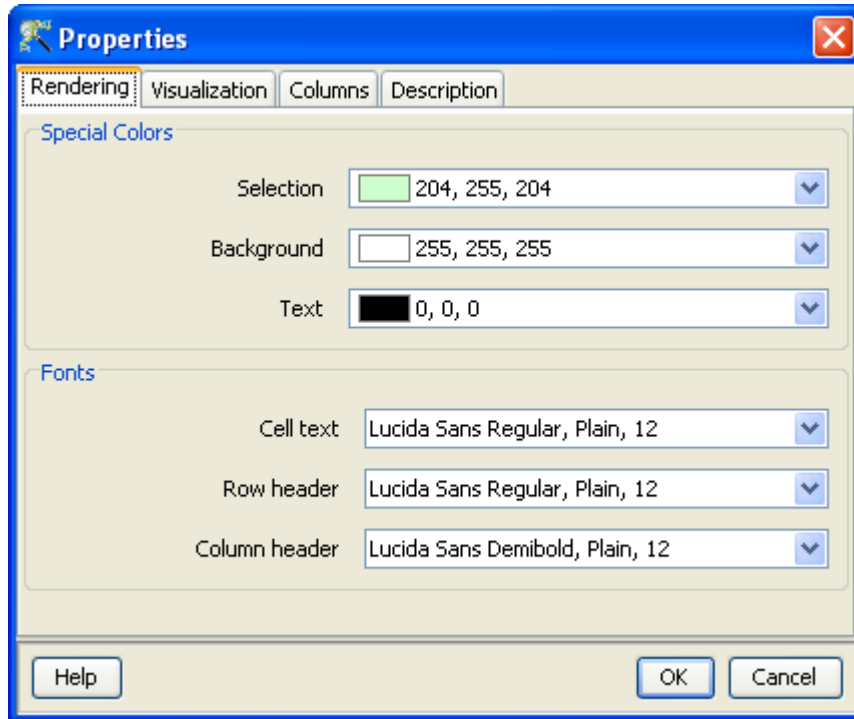


Figure 5.29: Summary Statistics Properties

Rendering: The rendering tab of the Summary Statistics View dialog allows you to configure and customize the fonts and colors that appear in the Summary Statistics View view.

Special Colors: All the colors in the Table can be modified and configured. You can change the Selection color, the Double Selection color, Missing Value cell color and the Background color in the table view. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the properties dialog. To change a color, click on the appropriate color bar. This will pop-up a Color Chooser. Select the desired color and click OK. This will change the corresponding color in the Table.

Fonts: Fonts that occur in the table can be formatted and configured. You can set the fonts for Cell text, row Header and Column Header. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a Font, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Visualization: The display precision of decimal values in columns, the row height and the missing value text, and the facility to enable and disable sort are configured and customized by options in this tab.

The visualization of the display precision of the numeric data in the table, the table cell size and the text for missing value can be configured. To change these, Right-Click on the table view and open the Properties dialog. Click on the visualization tab. This will open the Visualization panel.

To change the numeric precision. Click on the drop-down box and choose the desired precision. For decimal data columns, you can choose between full precision and one to four decimal places, or

representation in scientific notation. By default, full precision is displayed.

You can set the row height of the table, by entering a integer value in the text box and pressing Enter. This will change the row height in the table. By default the row height is set to 16.

You can enter any a text to show missing values. All missing values in the table will be represented by the entered value and missing values can be easily identified. By default all the missing value text is set to an empty string.

You can also enable and disable sorting on any column of the table by checking or unchecking the check box provided. By default, sort is enabled in the table. To sort the table on any column, click on the column header. This will sort the all rows of the table based on the values in the sort column. This will also mark the sorted column with an icon to denote the sorted column. The first click on the column header will sort the column in the ascending order, the second click on the column header will sort the column in the descending order, and clicking the sorted column the third time will reset the sort.

Columns: The order of the columns in the Summary Statistics View can be changed by changing the order in the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the Selected items to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the Selected items list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.

To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.

The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

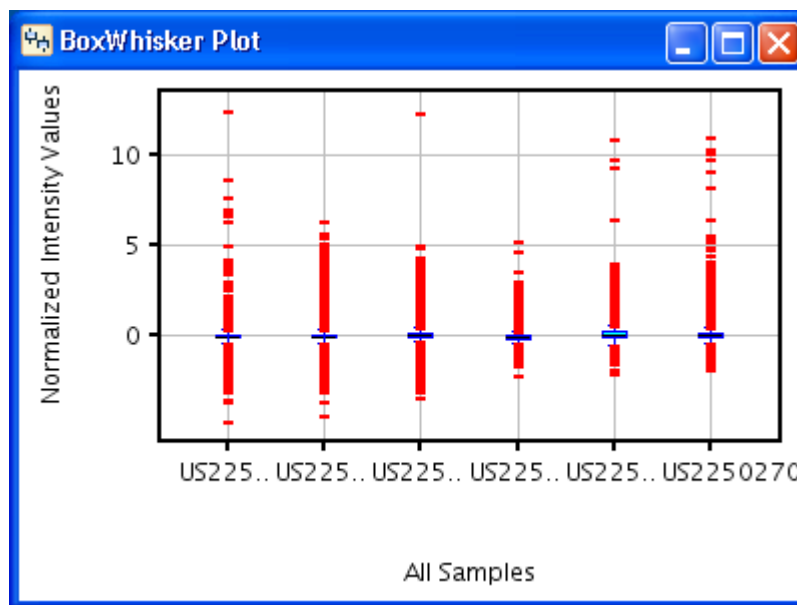



Figure 5.30: Box Whisker Plot

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.11 The Box Whisker Plot

The Box Whisker Plot is launched from View menu on the main menu bar with the active interpretation and the active entity list in the experiment. Alternately, Left-Click on the tool bar 'BoxWhisker'  icon will bring up the boxwhisker plot. The Box Whisker Plot presents the distribution of the of the conditions in the active interpretation with respect to the active entity list in the experiment. The box whisker shows the median in the middle of the box, the 25th percentile and the 75th percentile, or the 1st and 3rd quartile. The whiskers are extensions of the box, snapped to the point within 1.5 times the interquartile. The points

outside the whiskers are plotted as they are, but in a different color and could normally be considered the outliers. See Figure [5.30](#)

If the active interpretation is the default *All Samples* interpretation, the box whisker plot the distribution of each sample with respect to the active entity list. If an averaged interpretation is the active interpretation, the box whisker plot shows the distribution of the conditions in the averaged interpretation with respect to the active entity list. The legend window displays the interpretation on which the box whisker plot was launched.

Clicking on another entity list in the experiment will make that entity list active and the box whisker plot will dynamically display the current active entity list. Clicking on an entity list in another experiment will translate the entities in that entity list to the current experiment and display those entities in the box whisker plot.

The operations on the box whisker plot are similar to operations on all plots and will be discussed below. The box whisker plot can be customized and configured from the **Properties** dialog. If a columns are selected in the spreadsheet, the box whisker plot is be launched with the continuous columns in the selection. If no columns are selected, then the box whisker will be launched with all continuous columns in the active dataset.

5.11.1 Box Whisker Operations

The Box Whisker operations are accessed from the toolbar menu when the plot is the active window. These operations are also available by right-clicking on the canvas of the Box Whisker. Operations that are common to all views are detailed in the section [Common Operations on Plot Views](#). Box Whisker specific operations and properties are discussed below.

Selection Mode: The Selection on the Box Whisker plot is confined to only one column of plot. This is so because the box whisker plot contains box whiskers for many columns and each of them contain all the rows in the active dataset. Thus selection has to be confined to only to one column in the plot. The Box Whisker only supports the selection mode. Thus, left-clicking and dragging the mouse over the box whisker plot confines the selection box to only one column. The points in this selection box are highlighted in the density plot of that particular column and are also lassoed highlighted in the density plot of all other columns. Left-clicking and dragging, and shift-left-clicking and dragging selects elements and Ctrl-Left-Click toggles selection like in any other plot and appends to the selected set of elements.

5.11.2 Box Whisker Properties

The Box Whisker Plot offers a wide variety of customization and configuration of the plot from the Properties dialog. These customizations appear in three different tabs on the Properties window, labelled

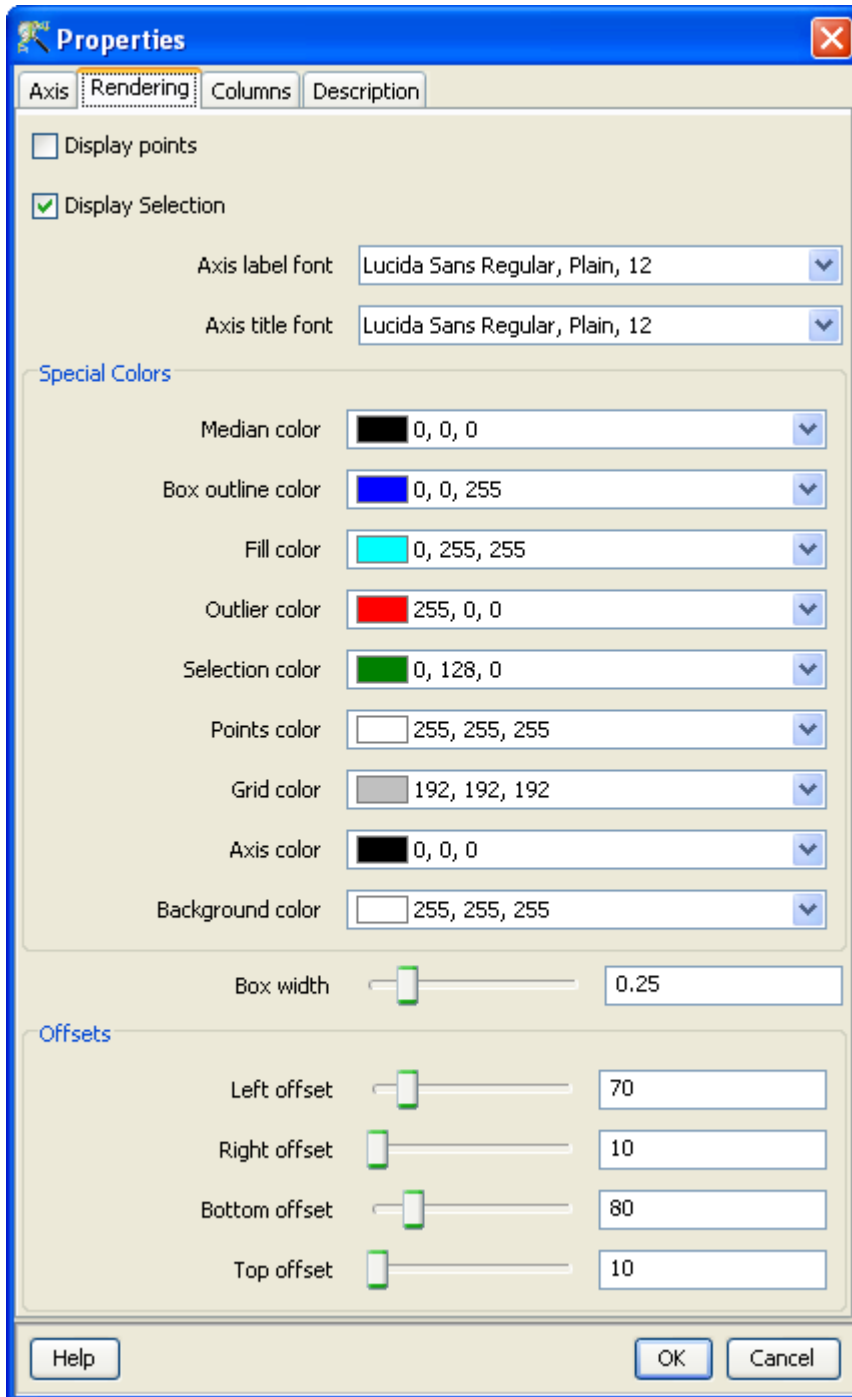


Figure 5.31: Box Whisker Properties

Axis, Rendering, Columns, and Description. See Figure 5.31

Axis: The grids, axes labels, and the axis ticks of the plots can be configured and modified. To modify these, Right-Click on the view, and open the Properties dialog. Click on the Axis tab. This will open the axis dialog.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For continuous data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Rendering: The Box Whisker Plot allows all aspects of the view to be configured including fonts, the colors, the offsets, etc.

Show Selection Image: The Show Selection Image, shows the density of points for each column of the box whisker plot. This is used for selection of points. For large datasets and for many columns this may take a lot of resources. You can choose to remove the density plot next to each box whisker by unchecking the check box provided.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the *Properties* dialog. To change a *Font*, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors on the box whisker can be configured and customized.

All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a *Color Chooser*. Select the desired color and click *OK*. This will change the corresponding color in the View.

Box Width: The box width of the box whisker plots can be changed by moving the slider provided. The default is set to 0.25 of the width provided to each column of the box whisker plot.

Offsets: The bottom offset, top offset, left offset, and right offset of the plot can be modified and configured. These offsets may be need to be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of the plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Columns: The columns drawn in the Box Whisker Plot and the order of columns in the Box whisker Plot can be changed from the Columns tab in the Properties Dialog.

The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the *Selected items* to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the *Selected items* list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.

To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.


The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.12 The Venn Diagram

The *Venn Diagram* is a special view that is used for capturing commonalities between entity lists, even across experiments. In **GeneSpring GX**, the user can choose entity lists from not only the active experiment but also from other experiments in a project. This is enabled by performing translation on the fly. This can happen provided homogene data exists for all the organisms of the selected experiments and the Entrez ID column exists for all the technologies of the selected entity lists.

The Venn Diagram is launched from the *View* menu on the main menu bar or from the tool bar. Alternately, Left-Click on the tool bar 'Venn Diagram'  icon will bring up the window to choose entity lists for the Venn diagram. Choose entity lists and click **Ok**; this will launch the Venn diagram with the chosen entity lists as circles of the Venn diagram after performing translation on the fly, if required. Venn diagram can be launched with a minimum of two and a maximum of three entity lists. See Figure 5.32


5.12.1 Venn Diagram Operations

Drag and drop operations on Venn diagram

After launching the Venn diagram, it is possible to add or replace an entity list, provided translation is possible between the chosen entity lists. From the navigator, choose an entity list and drag it into the Venn diagram view using the mouse. Drop outside the circles to add this entity list into the Venn diagram. Note that addition is possible only if the original venn diagram had two entity lists. Drop inside the non-overlapping part of a circle in the Venn diagram to replace that entity list with the chosen one.

While dragging an entity list into the Venn diagram, the mouse would indicate if addition/replacement is possible or not, at any position.

Selection from overlapping and unique regions

From the Venn diagram, select any region (overlapping or unique/non-overlapping part of the circles) with a left mouse click ; click on 'Create entity list from Selection'  icon. This will bring up a 2 step wizard titled "Create New Entity List".

- Step 1 of 2: The entity lists corresponding to the selected region are shown here along with their list associated values/columns. For each of the entity list, all or a subset of the list associated values/columns can be chosen. Some of the entity lists may be grayed out depending on the selected region from the Venn diagram.
- Step 2 of 2: The results including the selected entity lists and their list associated values/columns and corresponding annotations are displayed here. A *Find* functionality allows user to locate entity

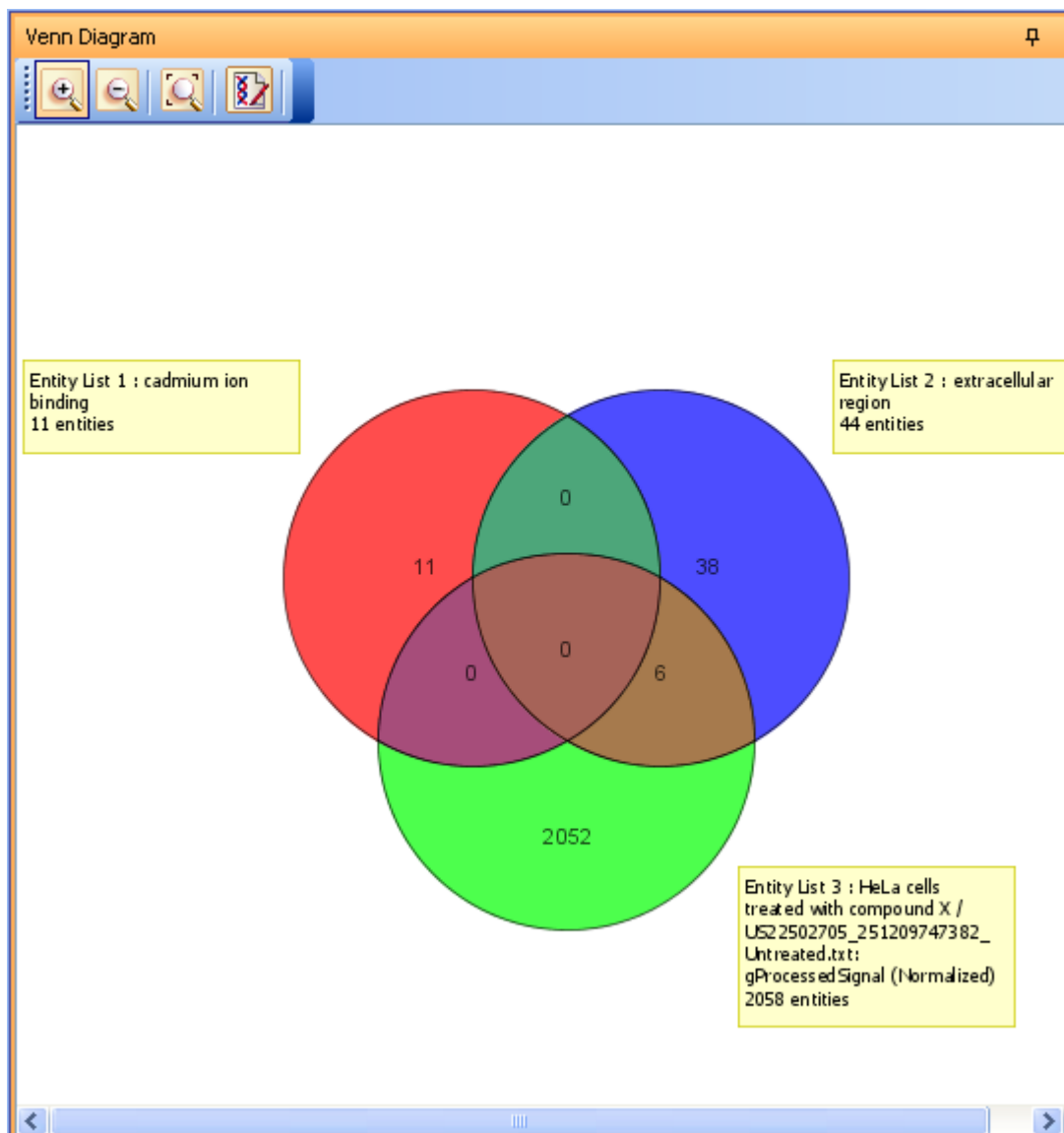


Figure 5.32: The Venn Diagram

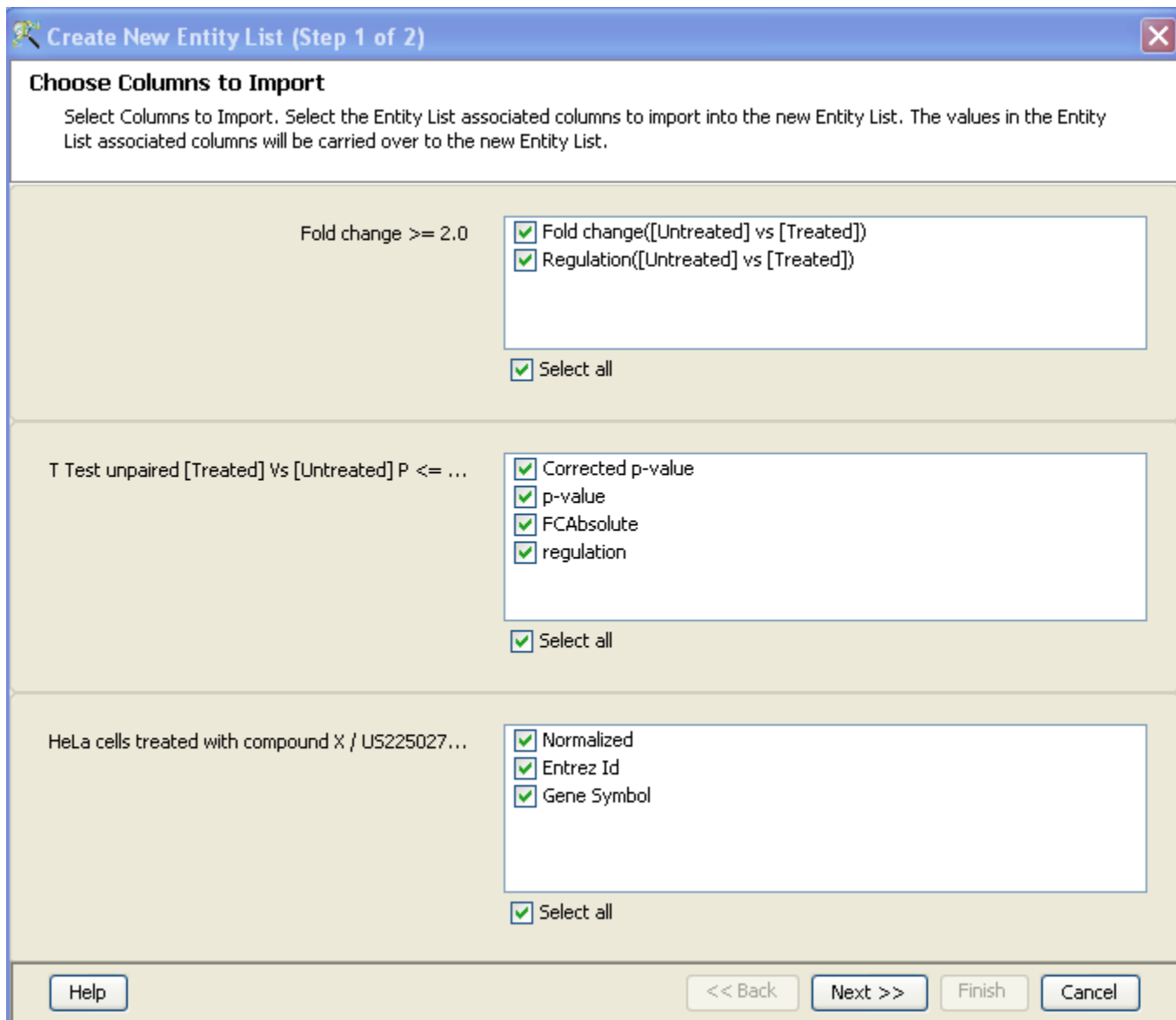


Figure 5.33: Create New Entity List from Venn Diagram

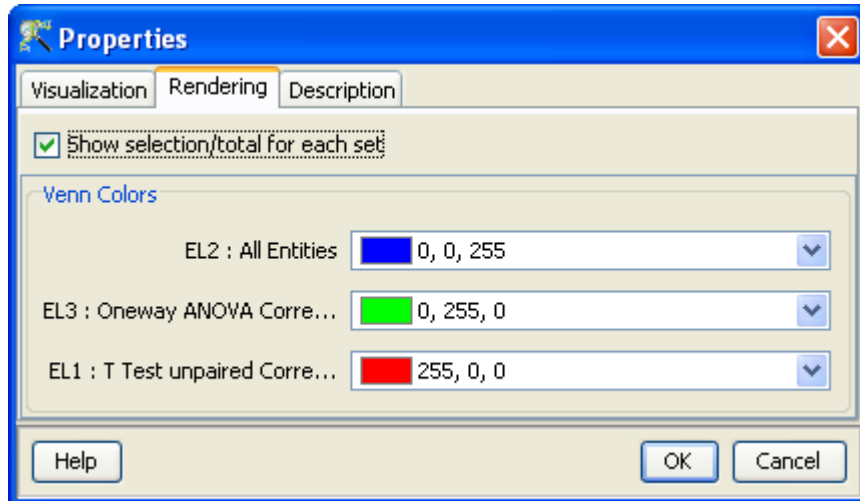


Figure 5.34: The Venn Diagram Properties

lists containing certain list associated values or annotations. It is possible to configure the columns by clicking the *Configure Columns* button. *Finish* exits the wizard after adding the newly created entity lists along with the chosen columns into the original experiment as a child node.

NOTE: *'Create New Entity List'* from the Venn diagram allows creation of new entity lists with the union of associated values/columns from the participating entity lists.

Right click operations

The operations on Venn diagram are accessible from the Right-Click menu on the Venn diagram. These operations are similar to the menu available on any plot. The Venn diagram is a lassoed view. Thus you can select any area within the Venn diagram. This will be shown with a yellow border and the genes in any in this area will be lassoed all across the project. Further, if you select any genes or rows from any other view, the Venn diagram will show the number of genes that in each area that are selected to the total number of genes in the area.

5.12.2 Venn Diagram Properties

The properties of the Venn diagram is accessible by Right-Click on the Venn diagram. See Figure 5.34

Visualization: The Venn diagram is drawn with chosen entity lists, either two or three. The visualization tab allows user to view the Venn diagram with all the permutation and combinations of the chosen

entity lists. For instance, if the Venn diagram was launched with entity lists E1, E2 and E3, from the visualization tab, user can choose to view the venn diagram with only E1 and E2 or E1 and E3 E2 and E3.

Rendering: The rendering tab of the Venn diagram properties dialog allows you to configure and customize the colors of the different entity list shown displayed in the Venn diagram.

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

5.13 LD Plot

You can launch the results of an LD Analysis from the experiment navigator. If the active entity list has entities from multiple Chromosomes then separate LD Plot nodes are created for each Chromosome.

The following steps guide you to launch an LD Plot:

- Click on an LD Plot node in the experiment navigator to launch the LD Plot.
- Select an option from the LD Measure menu: r^2 or $D - prime$ (default option).
- Drag the mouse pointer over the plot to select blocks of interest (Figure 27.4).
- Click on the "Create entity list from selection" icon in the toolbar, and save the blocks as entity lists.
 - Select "Single Entitylist" to save the entities in all the selected blocks as a single entity list.
 - Select "Blockwise Entitylist" to save the entities in each block as a separate entity list.

Refer to [LD Analysis](#) section for information on LD Analysis.

5.13.1 LD Plot Toolbar

You can select "Zoom" or "Selection" modes, reset the zoom , or clear the selection from the toolbar.

Zoom Mode:

Select a block of SNPs to launch an LD Plot of the block.



Figure 5.35: LD Plot Toolbar

Selection Mode:

- Select the blocks of SNPs.
- Click on the "create entity list from selection" icon to save each selected block as a separate entity list or all the blocks as a single entity list.

5.13.2 LD Measure Options

GeneSpring GX provides two LD measure visualizations: r^2 and D-prime.

r^2 :

Plots the raw r^2 score for a given marker pair. The r^2 is a measure of linkage disequilibrium between two genetic markers. For SNPs that have not been separated by recombination or have the same allele frequencies (perfect LD), $r^2 = 1$. In such cases, the SNPs are said to be redundant. Lower r^2 values indicate less degree of LD.

One useful property of r^2 for association studies is that its inverse value, $1/r^2$, provides a practical estimate of the magnitude by which the sample size must be increased in a study design to detect association between the disease and a marker locus, when compared with the size required for detecting association with the susceptibility locus itself [42].

D-prime:

This is the default plot and is displayed when the user turns on the LD plot track. This track plots the raw D-prime score for a given marker pair. D-prime is a measure of linkage disequilibrium between two genetic markers. A value of D-prime = 1 (complete LD) indicates that two SNPs have not been separated by recombination, while values of D-prime < 1 (incomplete LD) indicate that the ancestral LD was disrupted during the history of the population [35].

Note: Only D-prime values near one are a reliable measure of LD extent; lower D-prime values are usually difficult to interpret as the magnitude of D-prime strongly depends on sample size.

Source: http://hapmap.ncbi.nlm.nih.gov/gbrowse_help.html#genotypes

5.13.3 LD Plot Properties

You can open the Properties dialog from the context (right-click) menu.

The Properties dialog has three tabs, viz., Visualization, Rendering, and Description.

Visualization:

You can select the plot label from the drop-down box, and adjust the label offset, label length, and the margins (left, right, bottom, and top) using the respective sliders.

Rendering:

You can configure the color range adjusting the sliders for the Minimum, Center, and Maximum values, and the respective Color Chooser box.

Note: D-Prime: ranges from -1 to +1 R-Square: ranges from 0 to +1
--

Description

You can add an appropriate Title and Description for the plot and click OK.

Export As:

- You can export the plot as an Image or HTML page from the context (right-click) menu option.
- You save the Image as a .tiff, .png, .jpg, .jpeg, or .bmp file.

5.14 Haplotypes view

In **GeneSpring GX** you can launch the Haplotypes view from the Haplotype Entity List Inspector.

The view launches a list with the following columns:

Probe set Id or Name:

Provides the Name (Illumina) or Probe set id (Affymetrix) of the first SNP in the Haplo block.

F-Statistics p-value:

Provides F-statistic p-value for each Haplo block.

Haplotypes:

Lists all the haplotypes for each Haplo block.

T Statistics p-value:

Provides t-statistic p-values for each haplotype.

5.14.1 Haplotypes Context Menu

You can perform common tabular operations using the context (right-click) menu options, which are listed hereunder:

Select All Rows:

Allows you to select all the rows from the list, and then export the view as an image or html file.

Invert Row Selection:

Allows you to invert the row selection, and then use the "Limit to Row Selection" option to launch the selected rows in the view.

Clear Row Selection:

Allows you to clear the existing row selection.

Limit to Row Selection:

Allows you to launch the list with only the selected rows.

Copy View:

Allows you to copy the view to the clipboard.

Print:

Allows you to launch the view in the web browser, which

Export As:

Allows you to Export the view as an Image or HTML file:

- **Image:** Exports the view as an image in .tiff, .bmp, .jpg, .jpeg, .png, or .gif formats.
- **HTML:** Exports the view as an HTML file.

Properties:

Allows you to add a Title and Description for the view.

5.15 Genome Browser

The **GeneSpring GX** genome browser allows the viewing of expression data imposed against the genomic features. For more details on the same, refer to the chapter on [Genome Browser](#)

5.16 Plot Options

5.16.1 Plot Log10/Linear Values

In **GeneSpring GX**, the data in the experiments are in log2 scale and the views are launched with the data in log2 scale. This option '*Plot Log10/Linear Values*' allows the user to view the scatter plot and the profile plot in log10 scale or in linear scale. The signal values for this plot can be chosen to be in raw or normalized form and the plot will be launched with the chosen interpretation.

On clicking *View* → *Plot Log10/Linear Values*, a window comes up with options to choose the interpretation, the type of signal values (raw or normalized), the scale (log 10 or linear), and the plot (scatter or profile plot).

The legend accompanying the resultant plot will show the chosen parameters with which the plot was launched. The plot operations and properties remain the same as with the regular scatter plot and profile plot and are described in sections , ,.

5.16.2 Plot List Associated Values

This option allows the user to visually inspect the data associated with two entity lists either as a scatter plot, a histogram or as a profile plot. The list associated values includes the columns obtained during analysis such as Fold Change, Significance Analysis etc. It does not include the annotation columns associated with the entity list.

On selecting this option, a window appears in which the user needs to provide the necessary inputs. The option to choose entity lists as well as an interpretation is provided. The user can also select the type of visualization from the options (Scatter Plot, Histogram and Profile Plot) provided in the drop down. See figures 5.37, 5.38 and 5.39. There is also an option to see either the raw or normalized signal values for the entities in the entity list in the view. See 5.36

On clicking on **OK**, the tool shows the view specified. By default, the X-axis is the second column of the first entity list and the Y-axis is the third column of the second entity list (the first column in both being the identifier). The user is provided with an option of all the list associated values in the dropdown for the axes and can choose as required.

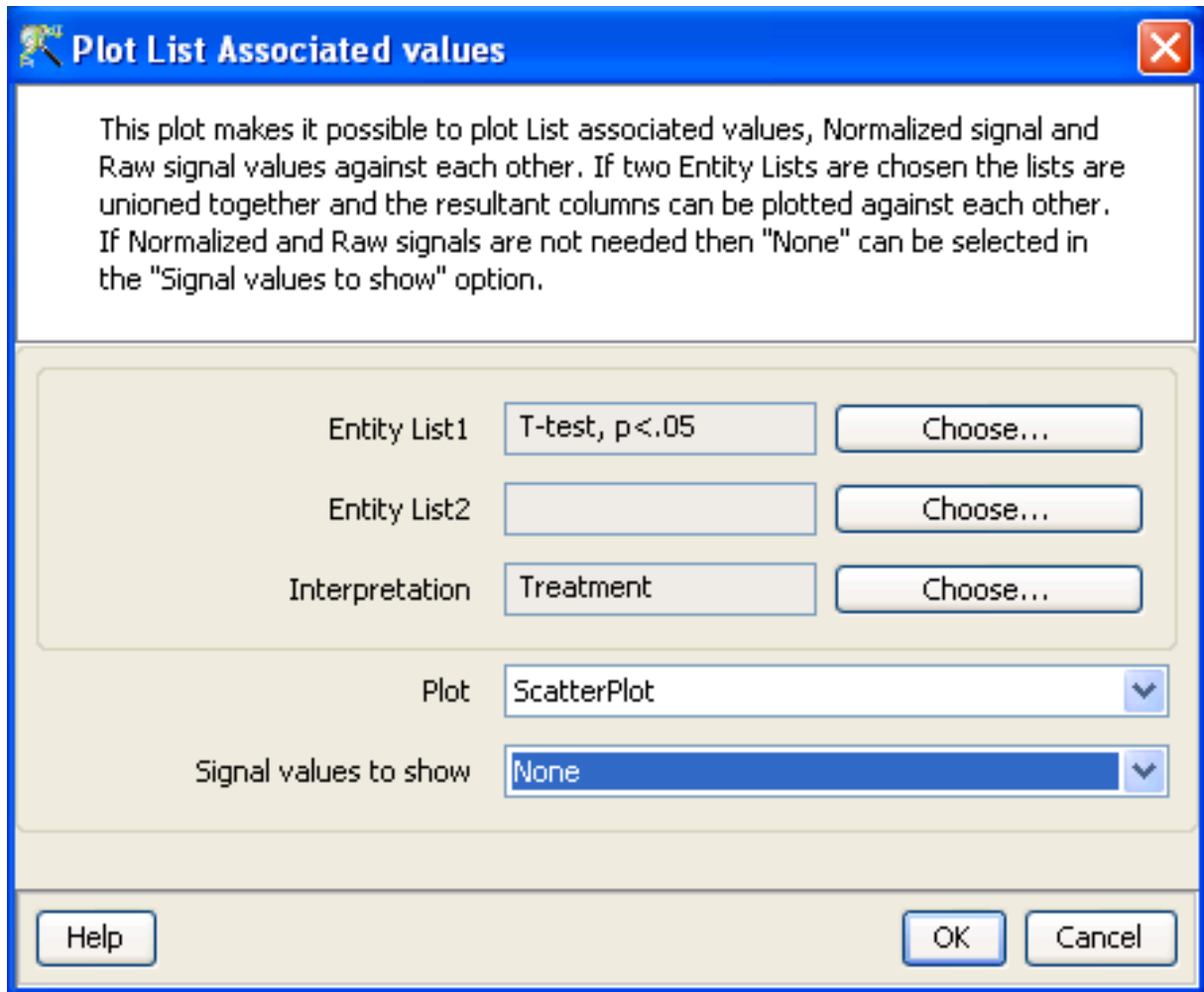


Figure 5.36: Plot List Associated Values

5.17 Miscellaneous operations

5.17.1 Save Current view

In **GeneSpring GX**, open views (Heatmaps, Classifications, Scatter Plot etc) are not saved in the experiment by default; so reopening or refreshing the experiment will not automatically bring up these views. Clicking on this option saves all currently open views in the experiment so these can be restored when the experiment is reopened.

There is one caveat though. Technology updates will not reflect in these saved views; so after a technology update, it is advisable to regenerate views which need updated annotation information.

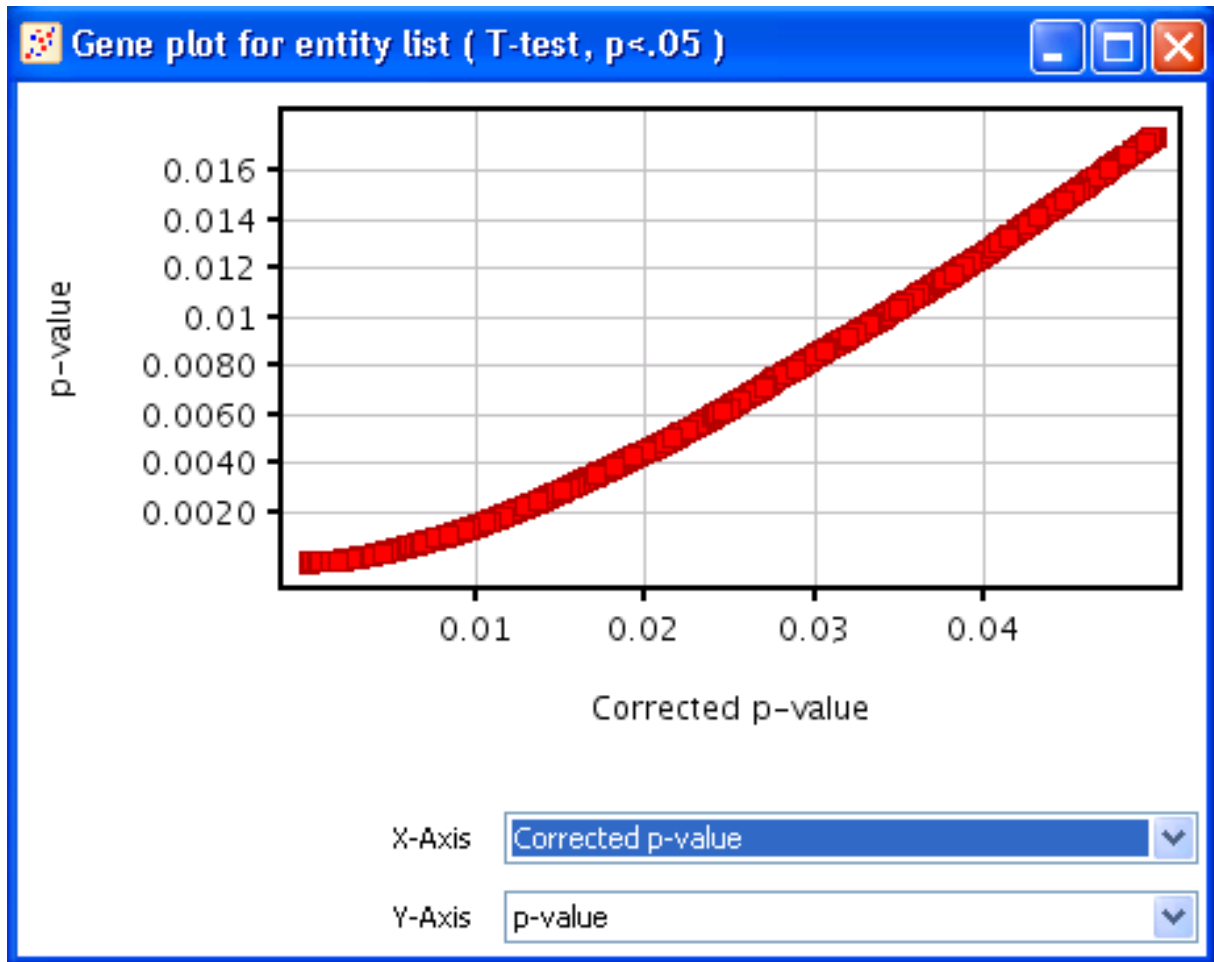


Figure 5.37: Plot List Associated Values-Scatter plot

5.17.2 Find Entity

Find Entity is a search functionality available through the *View* menu or with the key binding Ctrl-F. This brings up a window listing all entities. The *Find* tag at the bottom allows the user to input a string/value for the search. The other tags, *Find Next* and *Find Previous* select and highlight the next/previous entity that matches the search condition. *Select All* will select all entities that matches the search string entered in the *Find* text box. Configuration of columns can also be done through this step. Any selection here will reflect throughout the tool in all views.

5.17.3 Inspect Entities

Inspect Entities brings up the Entity inspector with the selected entities. Can also be called by the key binding Ctrl-I.

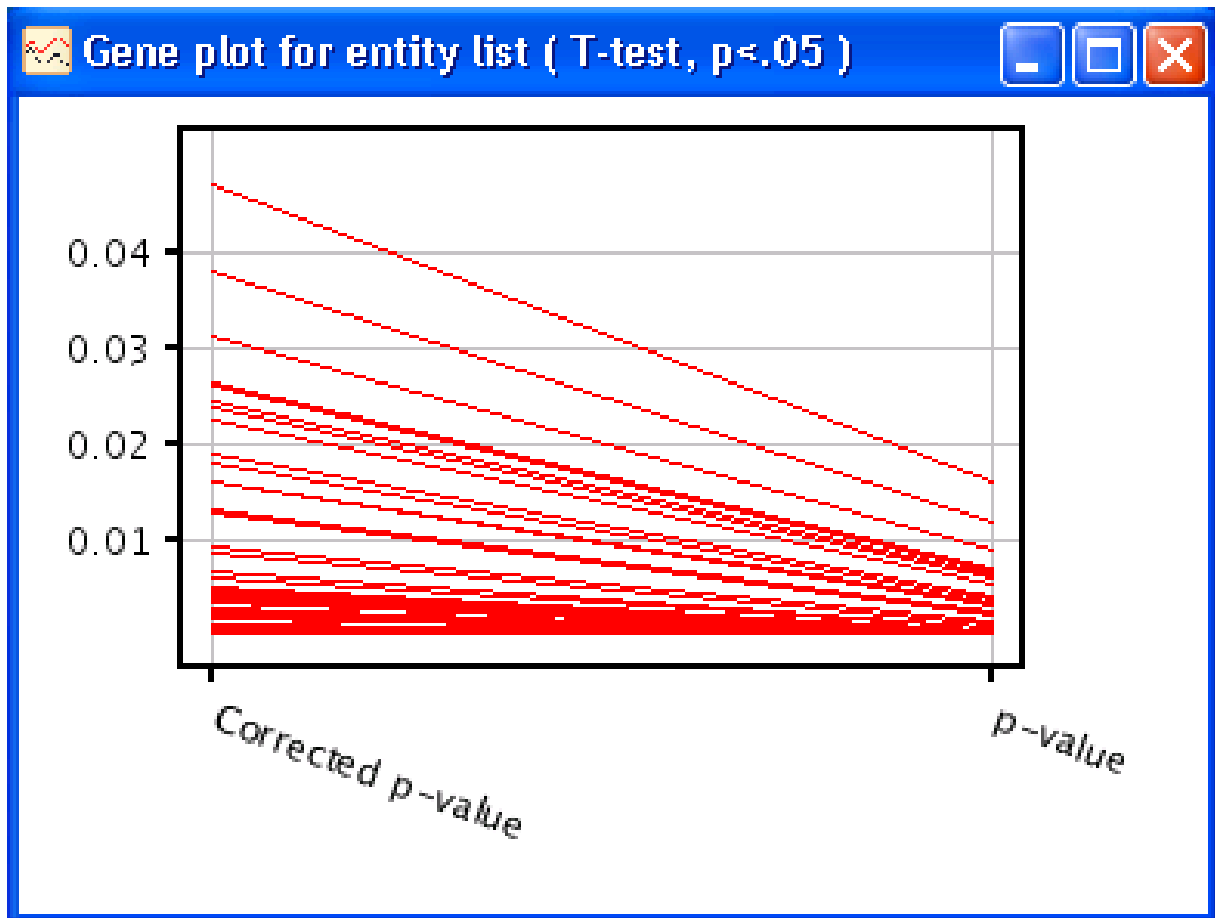


Figure 5.38: Plot List Associated Values-Profile plot

5.17.4 Properties

Properties or Ctrl-R brings up the properties windows relevant to the view on focus.

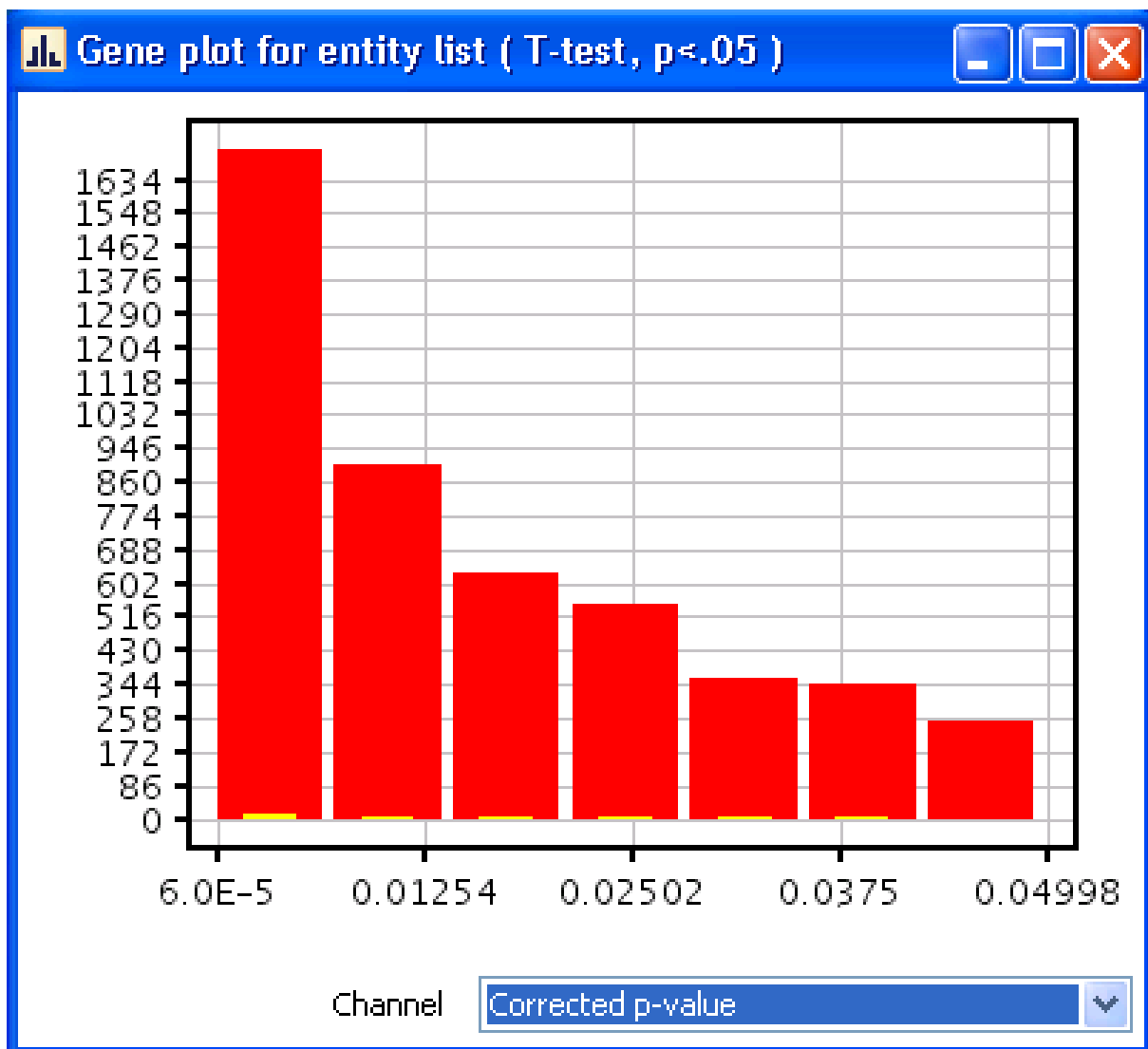


Figure 5.39: Plot List Associated Values-Histogram

Chapter 6

Analyzing Affymetrix Expression Data

GeneSpring GX supports the Affymetrix GeneChip technology. Most of the Affymetrix GeneChips can be analyzed using **GeneSpring GX**. To obtain a list of the supported chips, go to *Annotations* → *Create Technology* → *From Agilent Server*. This will display a list of supported chip types. Affymetrix technology can also be created if a custom CDF is being used. For more details refer to the section on [Affymetrix Technology creation using Custom CDF](#).

6.1 Running the Affymetrix Workflow

Upon launching **GeneSpring GX**, the startup is displayed with 3 options.

- **Create new project**
- **Open existing project**
- **Open recent project**

Either a new project can be created or a previously generated project can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on **OK** to proceed.

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment:** This allows the user to create a new experiment. (steps described below).

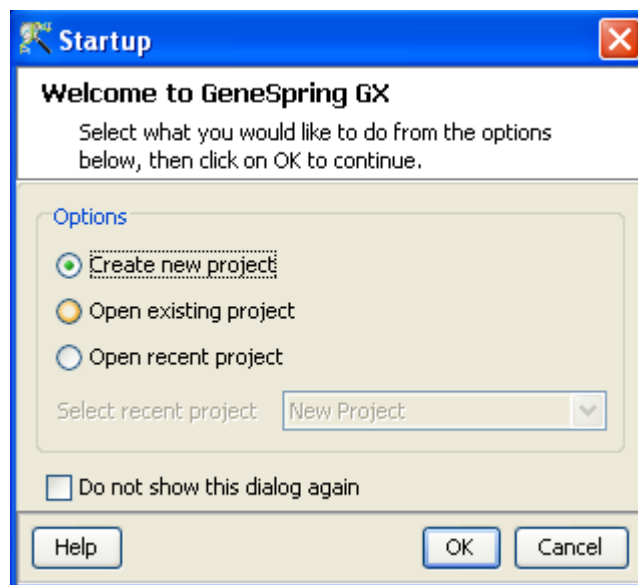


Figure 6.1: Welcome Screen

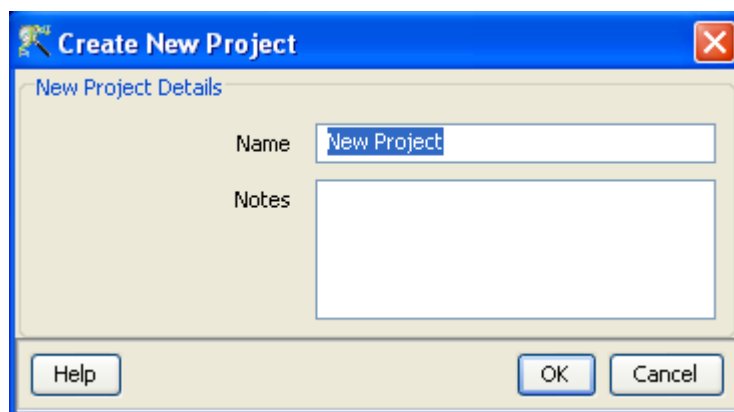


Figure 6.2: Create New project

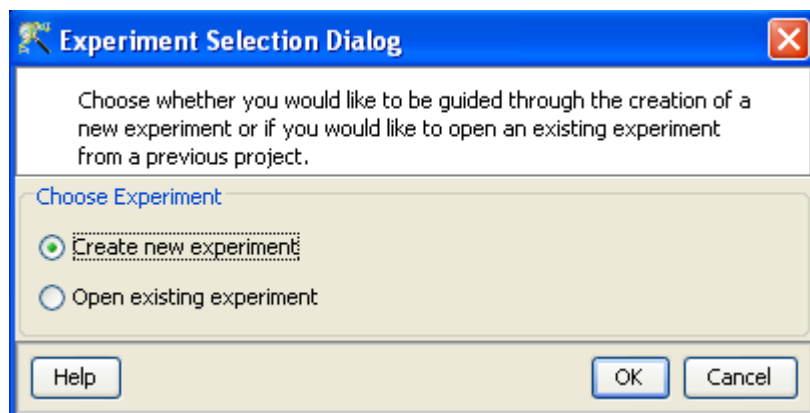


Figure 6.3: Experiment Selection

2. **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the experiment type gives the user the option to choose between the multiple experiment types namely Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Agilent miRNA, Generic Single Color, Generic Two Color, Pathway and RealTime-PCR experiment.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Guided Workflow**
2. **Advanced Analysis Workflow**

Guided Workflow is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Advanced Analysis**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples
3. Reorder

4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. On the other hand, a sample is created within **GeneSpring GX**, when it associates the data files with its appropriate technology (See the section on [Technology](#)). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the *Choose Samples* option. For selecting data files and creating an experiment, click on the *Choose File(s)* button, navigate to the appropriate folder and select the files of interest. Click on *OK* to proceed.

The technology specific for any chip type needs to be created or downloaded only once. Thus, upon creating an experiment of a specific chip type for the first time, **GeneSpring GX** prompts the user to download the technology from the update server. If an experiment has been created previously with the same technology, **GeneSpring GX** then directly proceeds with experiment creation. Clicking on the *Choose Samples* button, opens a sample search wizard, with the following search conditions:

1. **Search field:** Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition:** Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on *Add* button, or can be removed from the list using *Remove* button.

Files can either be removed or reordered during the data loading step using the *Remove* or *Reorder* button. Figures [6.4](#), [6.5](#), [6.6](#), [6.7](#) show the process of choosing experiment type, loading data, choosing samples and re-ordering the data files.

6.2 Data Processing

1. **File formats:** The data file should be present either as a CEL file or a CHP file or a TEXT file. However while creating an experiment; only one type of file (CEL/CHP.TXT) can be used.
2. **Raw signal values (CEL files):** In an Affymetrix Expression experiment, the term "raw" signal values refer to the linear data after summarization using a summarization algorithm (RMA, PLIER, GCRMA, LiWong and MAS5).

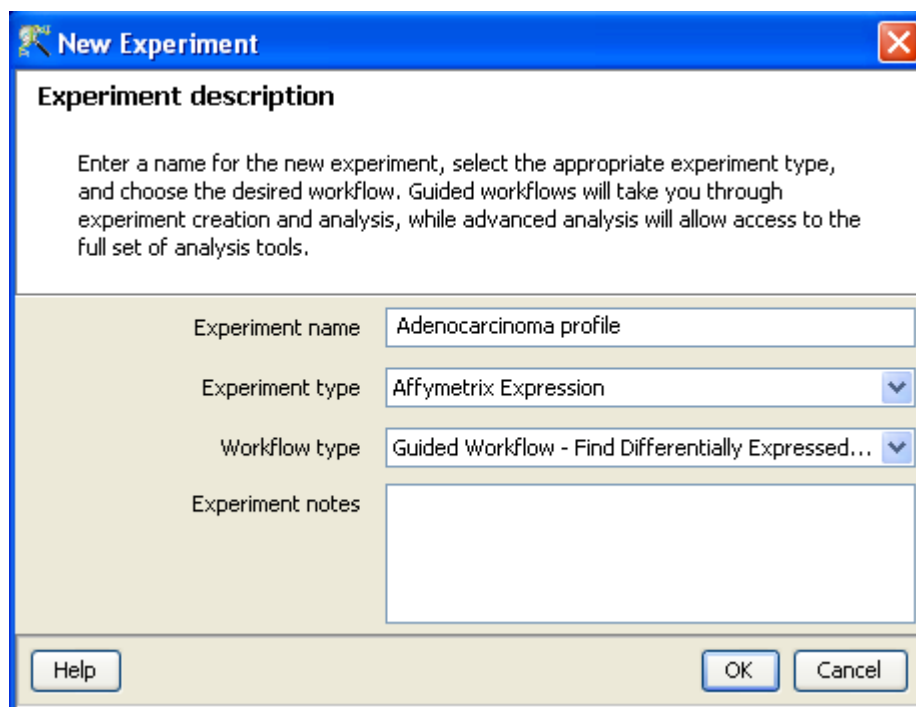


Figure 6.4: Experiment Description

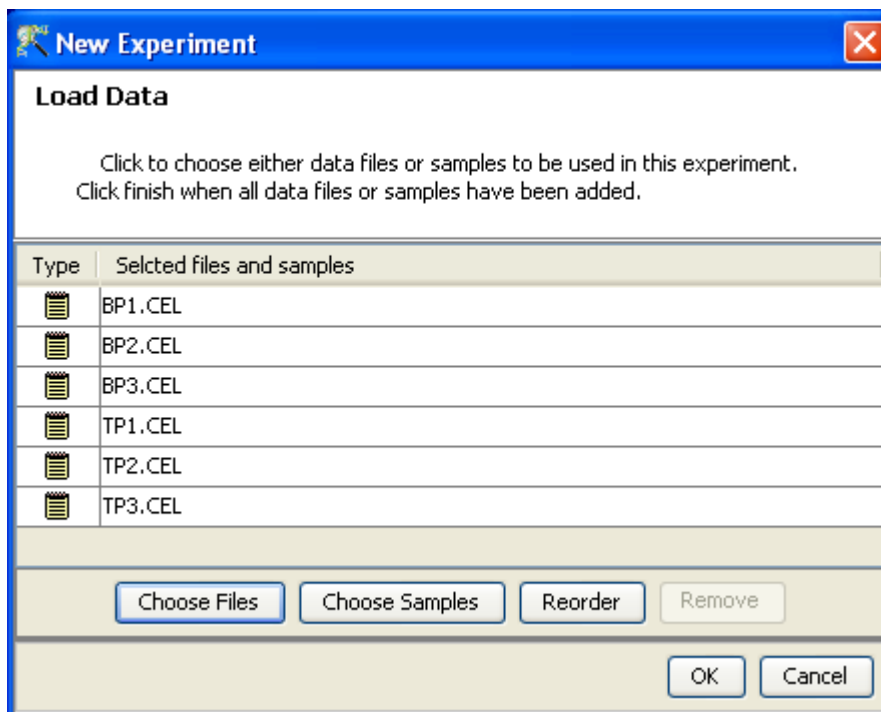


Figure 6.5: Load Data

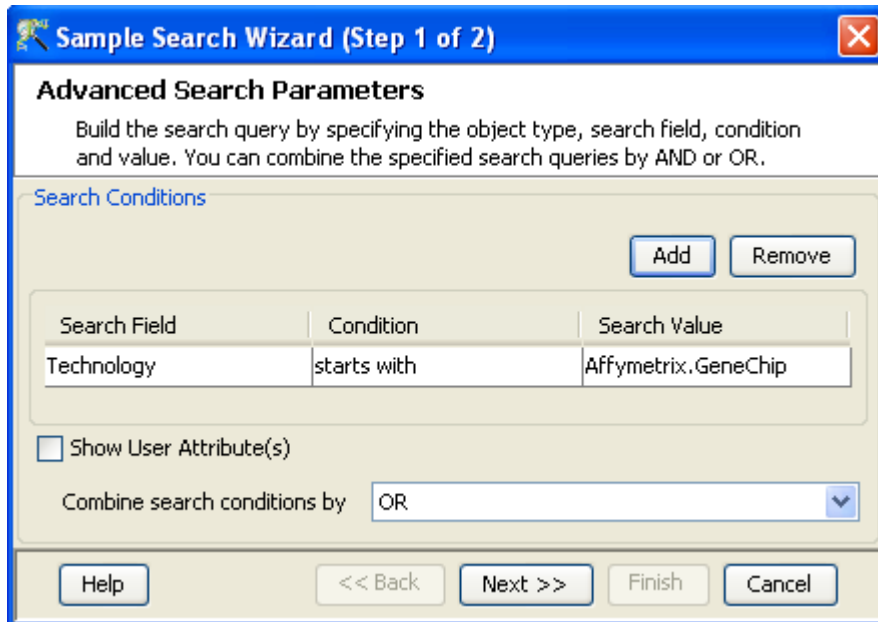


Figure 6.6: Choose Samples

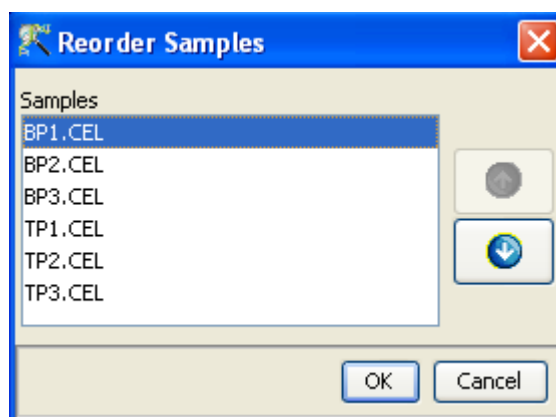


Figure 6.7: Reordering Samples

3. **Raw signal values (CHP files):** In an Affymetrix Expression experiment, the term "raw" files refers to the linear data obtained from the CHP files. In an Affymetrix Expression experiment, **GeneSpring GX** does not handle input data from CHP files if they are present in the log scale as the tool assumes that the data is in the linear scale and performs another log transformation.
4. **Normalized signal values (CEL files):** "Normalized" values are generated after log transformation and baseline transformation.
5. **Normalized signal values (CHP files):** The term "Normalized" refers to values generated after log transformation, normalization (Percentile Shift, Scale and Normalize to control genes) and baseline transformation.
6. **Treatment of on-chip replicates:** Not Applicable.
7. **Flag values:** The flag values are calculated only when MAS5 algorithm is chosen for summarization and is inclusive of the algorithm.
8. **Treatment of Control probes:** Not Applicable.
9. **Empty Cells:** Not Applicable.
10. **Sequence of events (CEL files):** The sequence of events involved in the processing of a CEL file is Summarization→log transformation→baseline transformation.
11. **Sequence of events (CHP files):** The sequences of events involved in the processing of a CHP file are log transformation→normalization→baseline transformation. If the data in the CHP file is already log transformed, then **GeneSpring GX** detects it and proceeds with the normalization step.
12. **Sequence of events (TXT files):** The sequences of events involved in the processing of a TXT file are log transformation→normalization→baseline transformation. The **GeneSpring GX** prompts the user to specify if the if the data in the Text file is already log transformed or not; User can then specify options for thresholding and normalization.

6.3 Guided Workflow steps

The *Guided Workflow* wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

Summary report (Step 1 of 8): The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized Expression values on the Y axis. An information message on the top of the wizard shows the sample processing details. By default, the *Guided Workflow* does RMA and baseline transformation to median of all samples. If the number of samples are more than 30, they are represented in a tabular column. On clicking the *Next* button it will proceed to the next step and on clicking *Finish*, an entity list will be created on which analysis can be done. By placing the cursor on the screen and

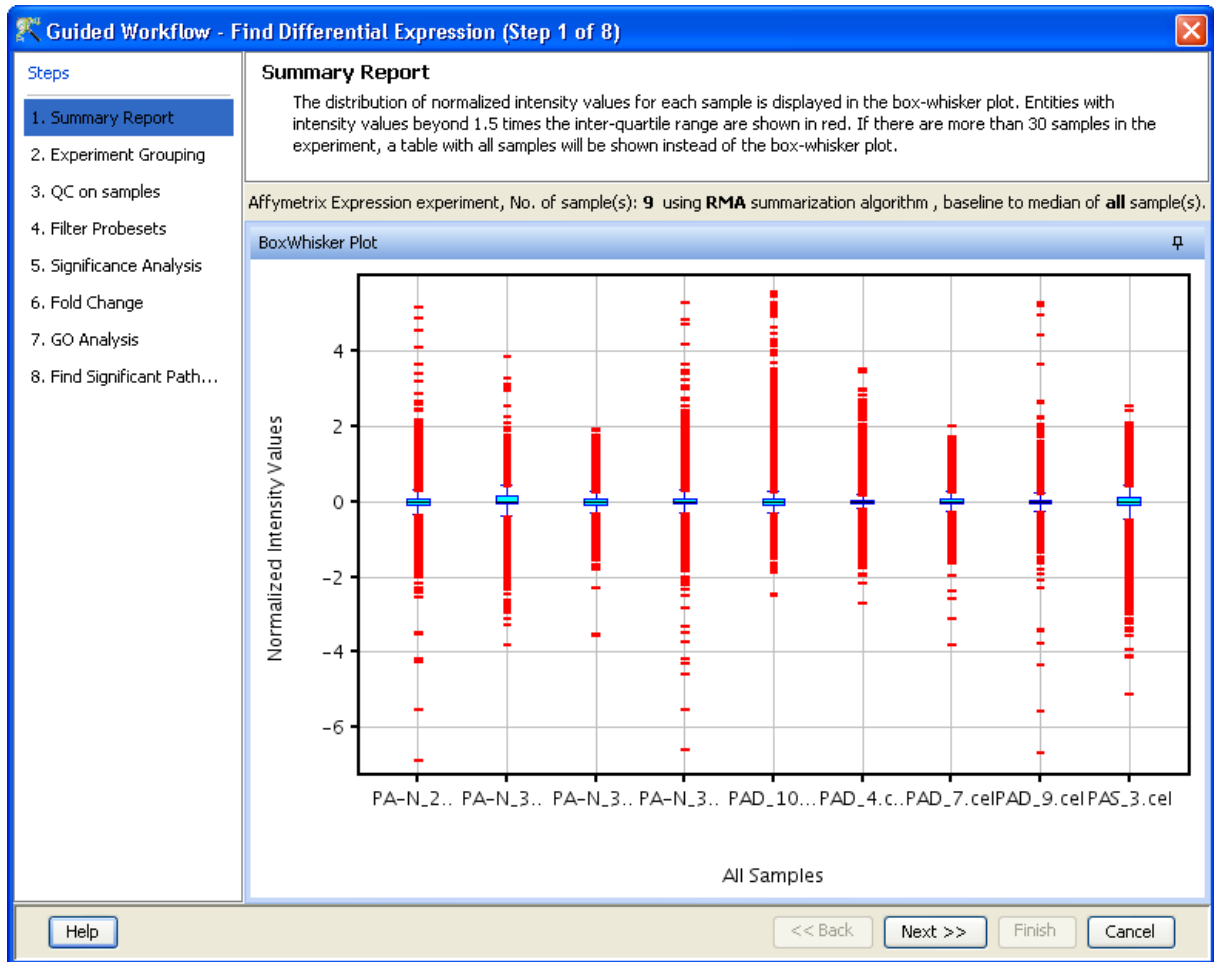




Figure 6.8: Summary Report

selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. On doing a right click, the options of invert selection is displayed and on clicking the same the selection is inverted i.e., all the probes except the selected ones are highlighted in green. Figure 6.8 shows the Summary report with box-whisker plot.

Note: In the *Guided Workflow*, these default parameters cannot be changed. To choose different parameters, use *Advanced Analysis*.

Experiment Grouping (Step 2 of 8): On clicking *Next*, the *Experiment Grouping* window appears which is the 2nd step in the **Guided Workflow**. It requires parameter values to be defined to group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the **Add parameter** button. Parameter values can be assigned by first selecting the desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on **Clear**. Press **OK** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Guided Workflow**. The other parameters can be used in the **Advanced Analysis**.





Note: The *Guided Workflow* does not proceed further without grouping information.

Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file  icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

```
A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50
```

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file  icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click → *Properties* → *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click → *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

Note: The *Guided Workflow* by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the guided wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures 6.9 and 6.10 respectively.

Quality Control on Samples (Step 3 of 8): The 3rd step in the *Guided Workflow* is the QC on samples which is displayed in the form of four tiled windows.

This window is disabled for TXT files.

- Internal controls and experiment grouping tabs

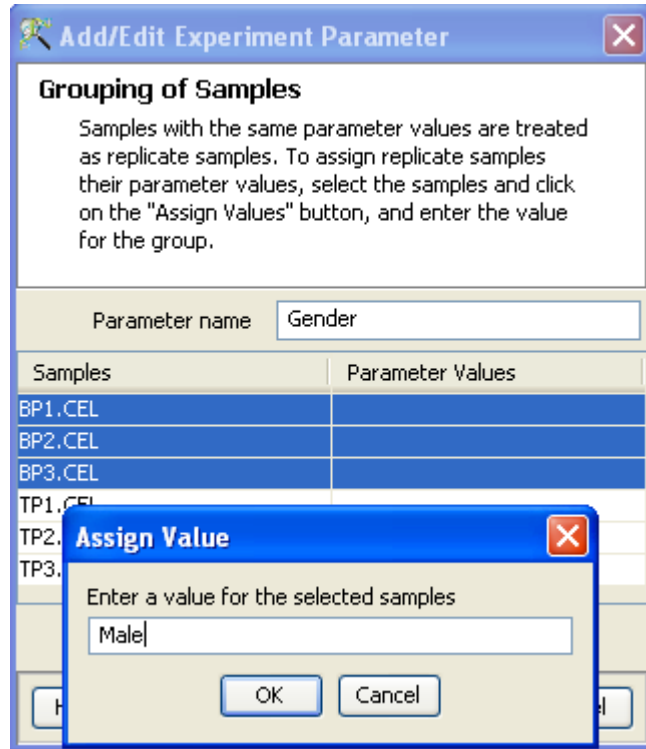


Figure 6.9: Experiment Grouping

- Hybridization controls
- PCA scores.
- Legend

QC generates four tiled windows as seen in Figure 6.11.

The views in these windows are lassoed i.e., selecting the sample in any of the view highlights the sample in all the views.

Internal Controls view shows RNA sample quality by showing 3'/5' ratios for a set of specific probesets which include the actin and GAPDH probesets. The 3'/5' ratio is output for each such probeset and for each array in the experiment. The ratios for actin and GAPDH should be no more than 3. A ratio of more than 3 indicates sample degradation and is shown in the table in red color. The *Experiment Grouping* tab, present in the same view shows the samples and the parameters assigned.

Hybridization Controls view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows

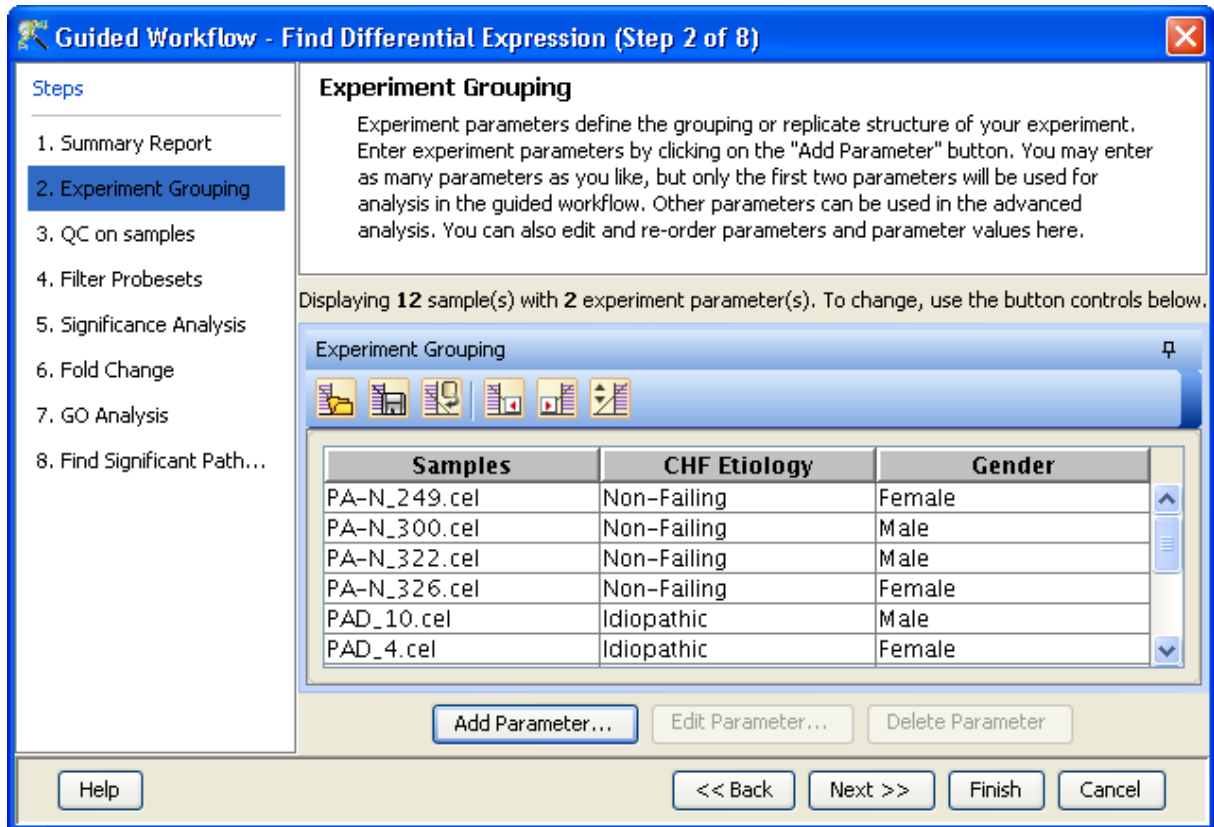


Figure 6.10: Edit or Delete of Parameters

viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove Samples* button allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed again on the newer sample set. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

Filter probesets (Step 4 of 8): This operation removes by default, the lowest **20 percentile** of all the intensity values and generates a profile plot of filtered entities. This operation is performed on the raw signal values. The plot is generated using the normalized (not raw) signal values and samples grouped by the active interpretation. The plot can be customized via the right-click menu. This filtered Entity List will be saved in the Navigator window. The Navigator window can be viewed after exiting from *Guided Workflow*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Annotations can be removed or added using *Configure Columns* button on the Entity Inspector. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The cutoff for filtering is set

QC on samples

Sample quality can be assessed by examining the values in the PCA plot and other experiment specific quality plots. To remove a sample from your experiment, select the sample from any of the views and click on the Add/Remove button. If a sample is removed, re-summation of the remaining samples will be performed.

Displaying 9 out of 9 samples retained in the analysis. To change, use the "Add/Remove Samples" button below.

Internal Controls: 3'/5' ratios

SampleN...	AFFX-HS...	AF
PA-N_24...	0.76208...	1.0
PA-N_30...	0.72959...	1.0
PA-N_32...	0.74363...	1.0
PA-N_32...	0.73684...	1.0
PAD_10.cel	0.76823...	1.0
PAD_4.cel	0.66818...	1.0
PAD_7.cel	0.66323...	1.0
PAD_9.cel	0.64563...	1.0

Experiment Grouping

Samples	New Pa
PA-N_24...	treated
PA-N_30...	treated
PA-N_32...	treated
PA-N_32...	treated
PAD_10.cel	untreat
PAD_4.cel	untreat
PAD_7.cel	untreat
PAD_9.cel	untreat

Hybridization Controls

3D PCA Scores

X Column: PCA Compo... Y Column: PCA Compo... Z Column: PCA Compo...

Add/Remove Samples

Figure 6.11: Quality Control on Samples

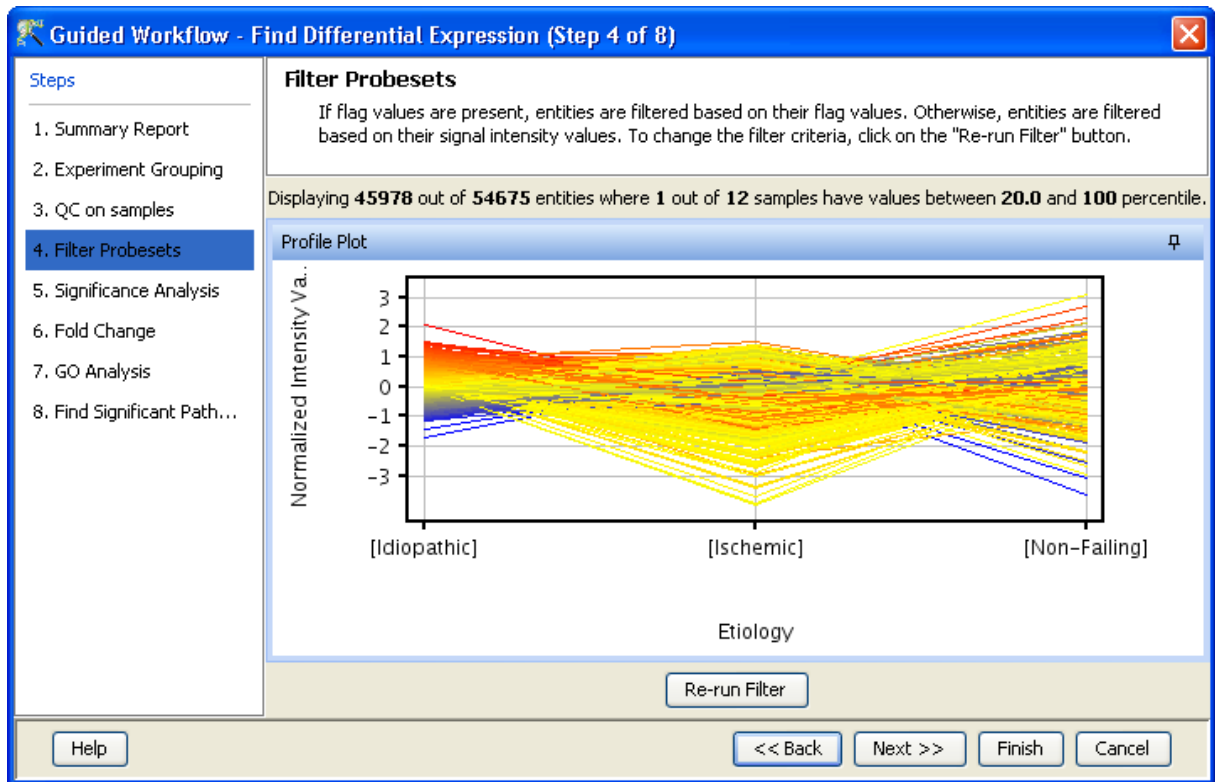


Figure 6.12: Filter Probesets-Single Parameter

at 20 percentile and which can be changed using the button *Rerun Filter*. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. Figures 6.12 and 6.13 are displaying the profile plot obtained in situations having single and two parameters.

Significance Analysis (Step 5 of 8): Depending upon the experimental grouping, **GeneSpring GX** performs either T-test or ANOVA. The tables below describe broadly the type of statistical test performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I*, has 2 groups, the normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 6.1: Sample Grouping and Significance Tests I

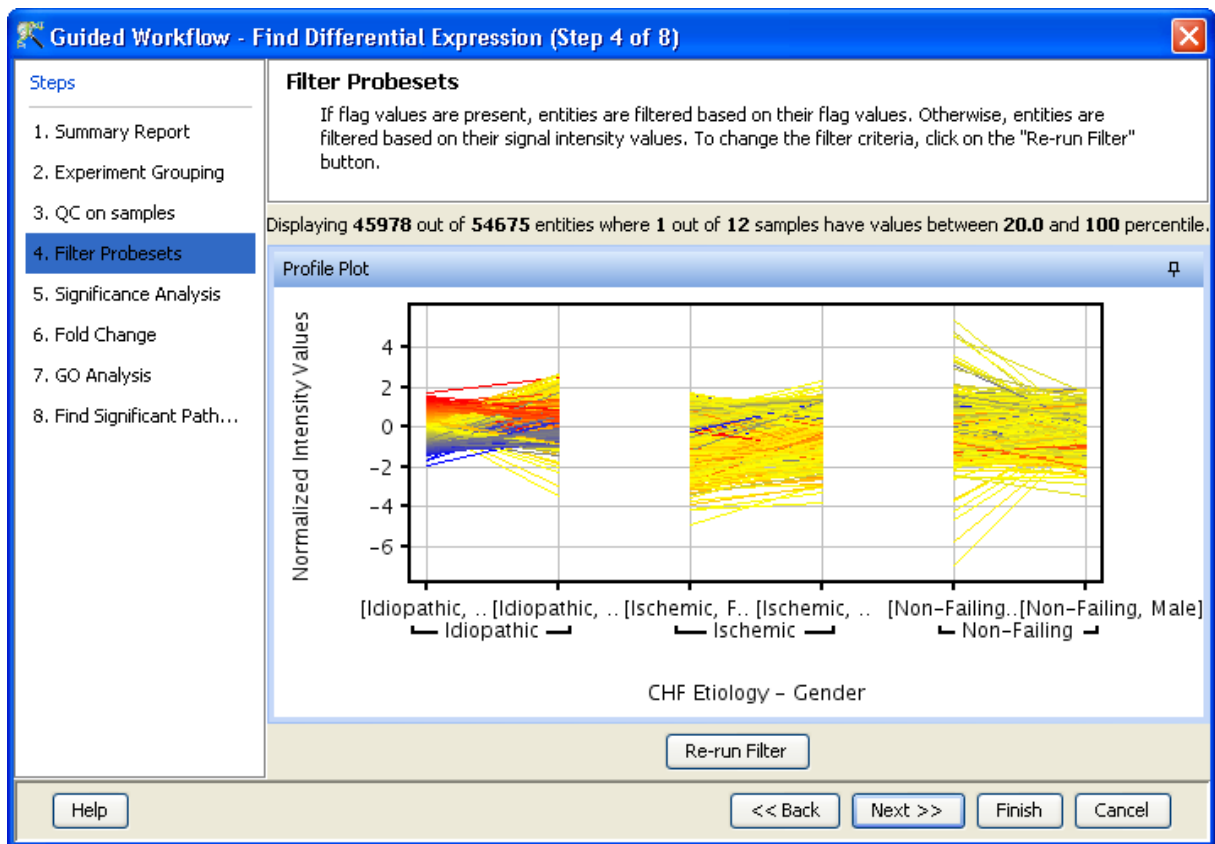


Figure 6.13: Filter Probesets-Two Parameters

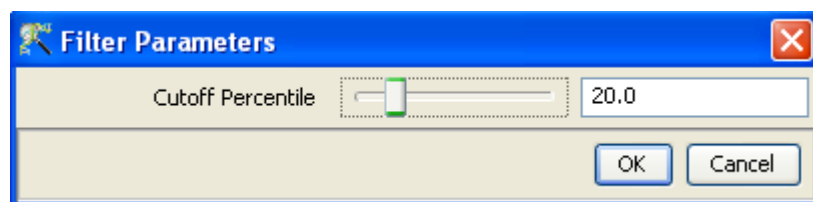


Figure 6.14: Rerun Filter

- **Example Sample Grouping II:** In this example, only one group, the tumor, is present. T-test against zero will be performed here.

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 6.2: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When 3 groups are present (normal, tumor1 and tumor2) and one of the groups (tumor2 in this case) does not have replicates, statistical analysis cannot be performed. However if the condition tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 6.3: Sample Grouping and Significance Tests III

- **Example Sample Grouping IV:** When there are 3 groups within an interpretation, One-way ANOVA will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 6.4: Sample Grouping and Significance Tests IV

- **Example Sample Grouping V:** This table shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.
- **Example Sample Grouping VI:** In this table, a two-way ANOVA will be performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 6.5: Sample Grouping and Significance Tests V

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 6.6: Sample Grouping and Significance Tests VI

- **Example Sample Grouping VII:** In the example below, a two-way ANOVA will be performed and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 6.7: Sample Grouping and Significance Tests VII

Statistical Tests: T-test and ANOVA

- **T-test: T-test unpaired** is chosen as a test of choice with a kind of experimental grouping shown in Table 1. Upon completion of T-test the results are displayed as three tiled windows.
 - A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *Regulation*.
 - *Differential expression analysis report* mentioning the Test description i.e. test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic*

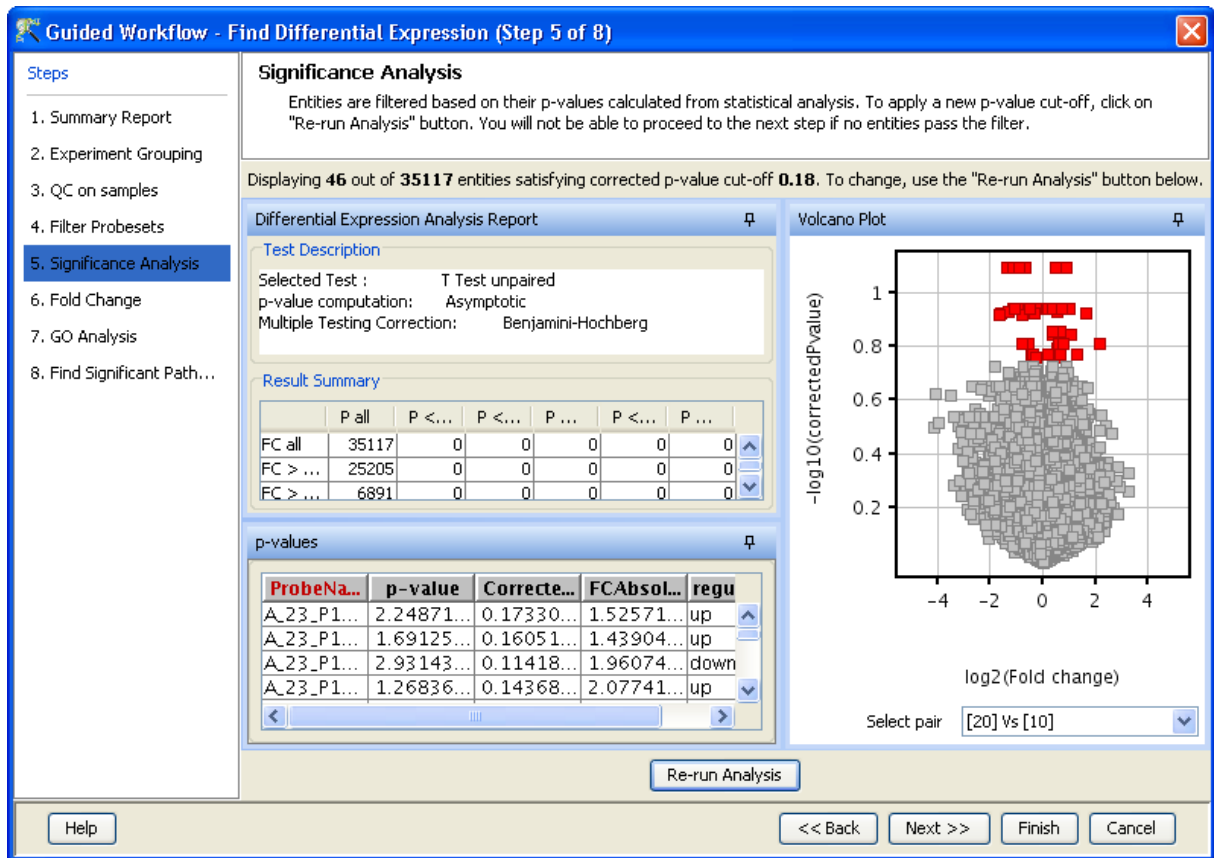


Figure 6.15: Significance Analysis-T Test

or *Permutative*).

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

- **Analysis of variance(ANOVA)**: ANOVA is chosen as a test of choice under the experimental grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:
 - A *p-value table* consisting of probe names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.
 - *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and p-value computation type (*Asymptotic* or *Permutative*).
 - *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

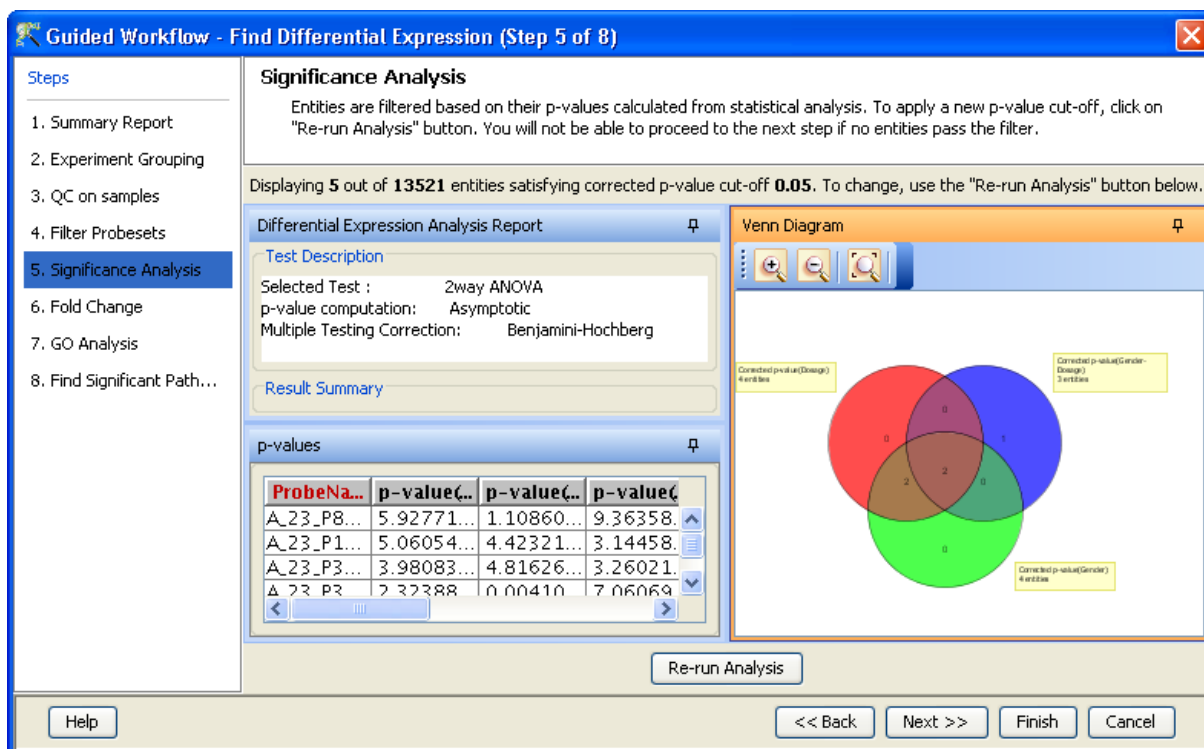


Figure 6.16: Significance Analysis-Anova

Special case: In situations when samples are not associated with at least one possible permutation of conditions (like Normal at 50 min and Tumor at 10 min mentioned above), no p-value can be computed and the **Guided Workflow** directly proceeds to **GO analysis**.

Fold-change (Step 6 of 8): **Fold change analysis** is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis. The wizard shows a table consisting of 3 columns: Probe Names, Fold change value and regulation (up or down). The regulation column depicts which one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using **Re-run Filter**. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1. A profile plot is also generated. Upregulated entities are shown in red. The color can be changed using the Right-click → *Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cutoff in the experiment Navigator.

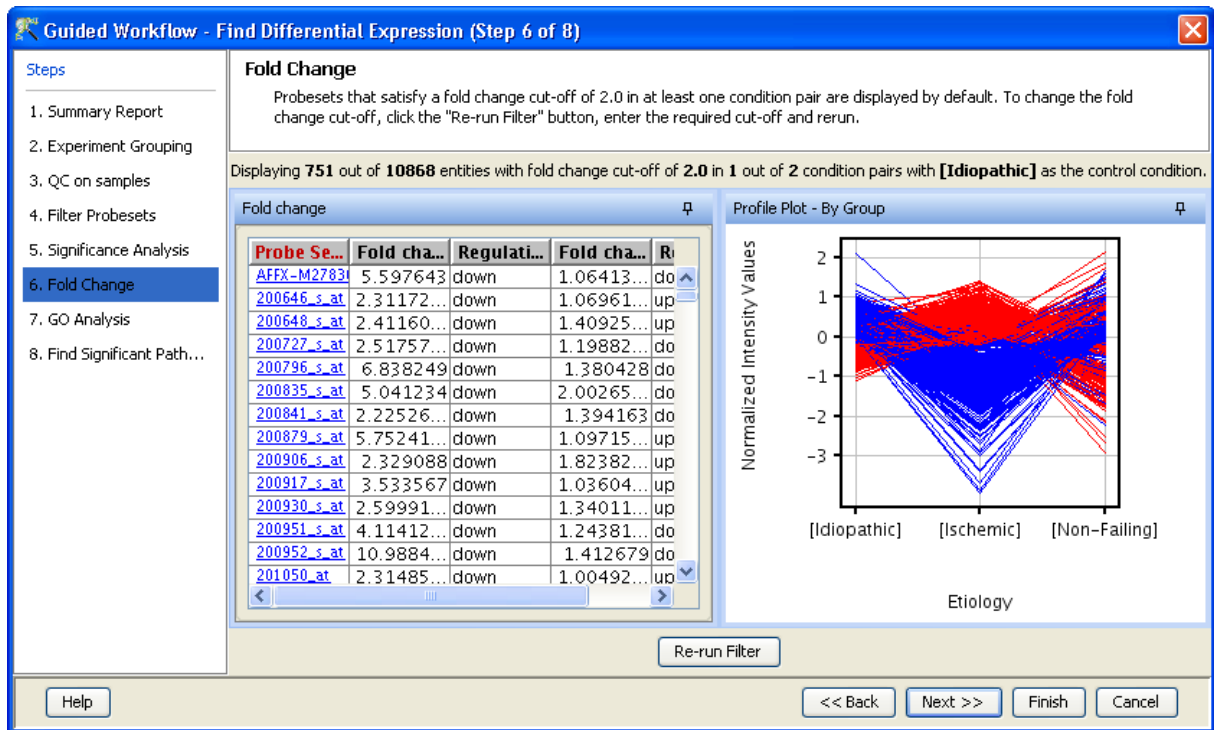


Figure 6.17: Fold Change

Note: Fold Change step is skipped and the *Guided Workflow* proceeds to the *GO Analysis* in case of experiments having 2 parameters.

Fold Change view with the spreadsheet and the profile plot is shown in Figure 6.17.

Gene Ontology(GO) Analysis (Step 7 of 8): The *GO Consortium* maintains a database of controlled vocabularies for the description of molecular function, biological process and cellular location of gene products. The GO terms are displayed in the Gene Ontology column with associated *Gene Ontology Accession* numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

The GO analysis wizard shows two tabs comprising of a spreadsheet and a *GO tree*. The *GO Spreadsheet* shows the *GO Accession* and *GO terms* of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However, the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.1 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

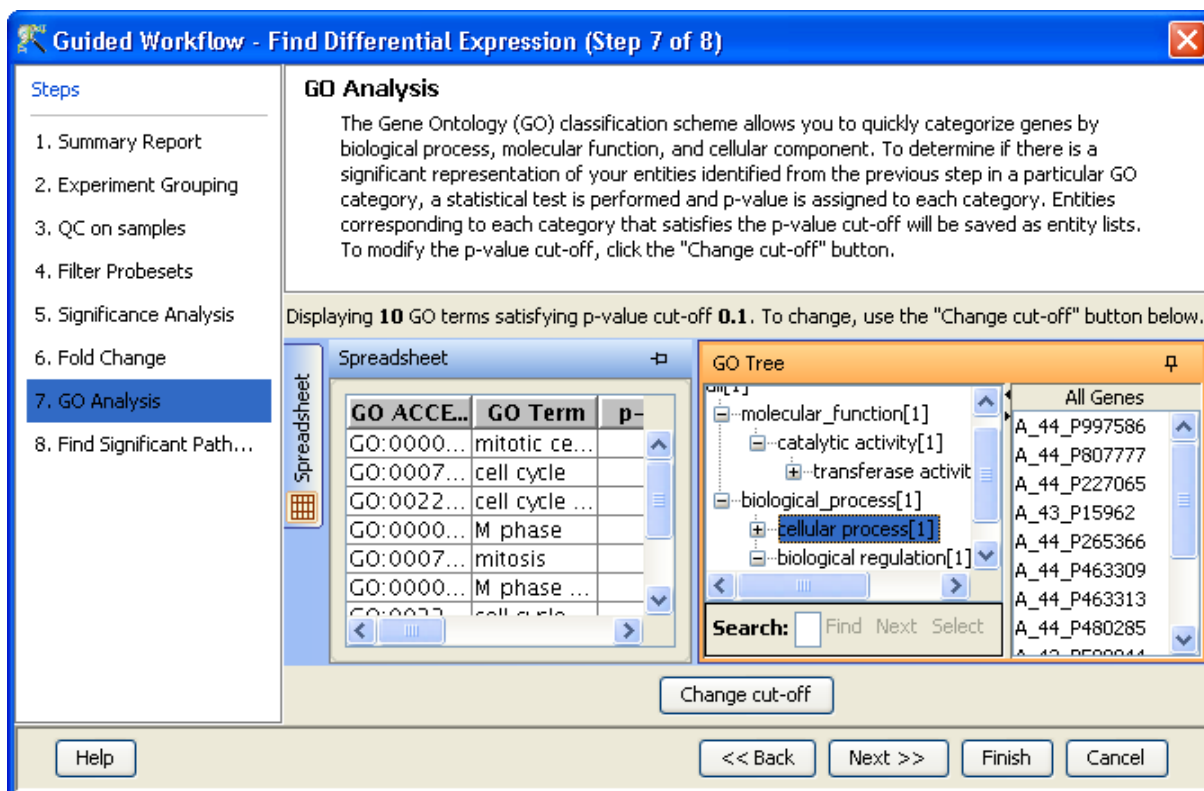


Figure 6.18: GO Analysis

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button is also provided on the GO tree panel to search using some keywords

Note : In **GeneSpring GX** GO analysis implementation, all the three component: Molecular Function, Biological Processes and Cellular location are considered together.

On finishing the GO analysis, the *Advanced Workflow* view appears and further analysis can be carried out by the user. At any step in the Guided workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any) and the *Advanced Workflow* view appears.

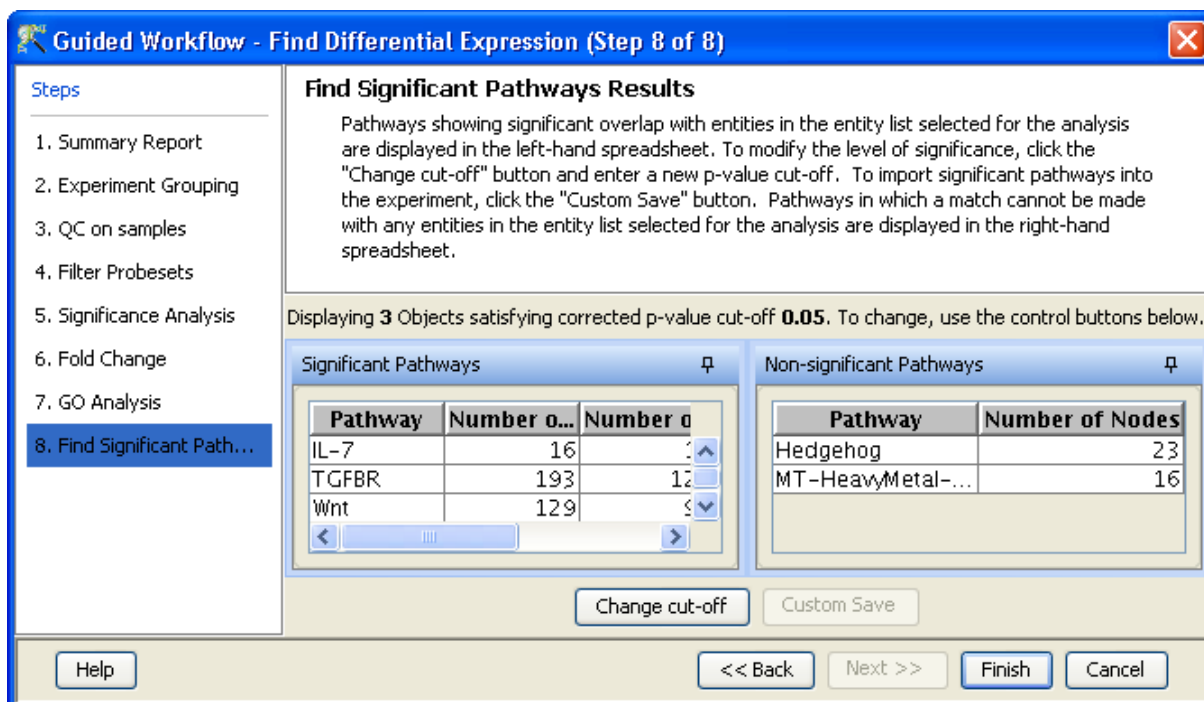


Figure 6.19: Find Significant Pathways

Find Significant Pathways (Step 8 of 8): This step in the Guided Workflow finds relevant pathways from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the fold change (step 6 of 8). This view shows two tables-

- The Significant Pathways table shows the names of the pathways as well as the number of nodes and entities in the pathway and corresponding p-values. It also shows the number of entities that are similar to the pathway and the entity list. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.
- The Non-significant Pathways table shows the pathways in the tool that do not have a single entity in common with the ones in the given entity list.

The user has an option of changing the p-value cut-off (using *Change cutoff*) and also to save specific pathways using the *Custom Save* option. See figure 6.19. On clicking, *Finish* the main tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of the Guided Workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in the Guided Workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any).

Note: In case the user is using **GeneSpring GX** for the first time, this option will give results using the demo pathways. The user can upload the pathways of his/her choice by using the option *Import BioPAX pathways* under **Tools** in the **Menu** bar. Later instead of reverting to the Guided Workflow the user can use the option *Find Significant Pathways* in **Results Interpretation** under the same Workflow.

The default parameters used in the *Guided Workflow* is summarized below

	Parameters	Parameter values
Expression Data Transformation	Thresholding	Not Applicable
	Normalization	Quantile
	Baseline Transformation	Median of all Samples
	Summarization	RMA
Filter by		
1.Flags	Flags Retained	Not Applicable
2.Expression Values	(i) Upper Percentile cutoff	100
	(ii) Lower Percentile cutoff	20.0
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1
Find Significant Pathways	p-value cutoff	0.05

Table 6.8: Table of Default parameters for Guided Workflow

6.4 Advanced Workflow

The **Advanced Workflow** offers a variety of choices to the user for the analysis.

- Several different summarization algorithms are available for probeset summarization.
- There are options for baseline transformation of the data and for creating different interpretations.
- Supports import of TXT files through templates.

To create and analyze an experiment using the **Advanced Workflow**, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Advanced. Clicking **OK** will open a New Experiment Wizard, which then proceeds as follows:

6.4.1 Creating an Affymetrix Expression Experiment

An **Advanced Workflow** analysis can be done using either CEL or CHP or TXT files. However, a combination of the file types are not allowed.

The following steps describe how to import a CEL/CHP/TXT file into **GeneSpring GX** .

Note that while importing text files, **GeneSpring GX** will automatically check with available templates and try to import based on a template. Standard files created in GCOS and Expression Console are available as templates in **GeneSpring GX** . There are two such templates available for each of GCOS and Expression Console - Metrics file where each sample is a file and Pivot file where multiple samples are in a file.

In order to create experiment from pivot files, GeneSpring assumes that the input files are in certain format. Column names ending with 'Signal' are treated as signal columns and column names ending with 'Detection' are treated as Flag columns. If the file contains column names like 'T1_Signal' and 'T1_Detection', then a sample 'T1' is created with 'T1_Signal' as signal column and 'T1_Detection' as flag column.

When a new TXT file is input, it is checked against these standard templates. If it matches any of these standard, it is imported based on that template. But if it does not match with any of these templates, then the user is taken through a custom template creation procedure. Templates created and saved by the user are added to the list of available templates, which then can be chosen as standard template while importing TXT files. Please refer to the section for details on custom template creation.

Step 1 of 10 : Load data As in case of **Guided Workflow**, either data files can be imported or else pre-created samples can be used.

- For loading new CEL/CHP/TXT files, use **Choose Files**.
- If the CEL/CHP/TXT files have been previously used in experiments **Choose Samples** can be used.

The **Load Data** window is shown in Figure 6.20.

Step 2 of 10: Choose Technology and Template This step comes up only for sample files in TXT format. The *Select Technology* drop down lists all the Affymetrix technologies available while *Choose a Template* shows available templates (those prepackaged in the tool and those saved by the user) as well as option to choose 'Custom Template'.

If 'Custom Template' is chosen, the user has to specify a sample data file to be considered as template and the number of samples in that data file. A 'Template name' also has to be input so that the custom template can be saved for future use. If a custom template is chosen, the wizard goes through steps 3 to 5, specific to creation of custom template. These steps are skipped if the TXT file is of the standard template form.

The **Choose Technology and Template** window is shown in Figure 6.21.

Skip the custom template creation steps and Go to [Step 6 of 10 : Select ARR Files](#)

Note that steps 3 to 5 of this wizard are applicable only when custom template has to be created to import non-standard TXT files.

Step 3 of 10 : Select Row Scope for Import This window shows the first few rows of the chosen template file (by default, it is 100 rows; can be changed from *Tools* → *Options menu*. The user can define the scope of the import here.

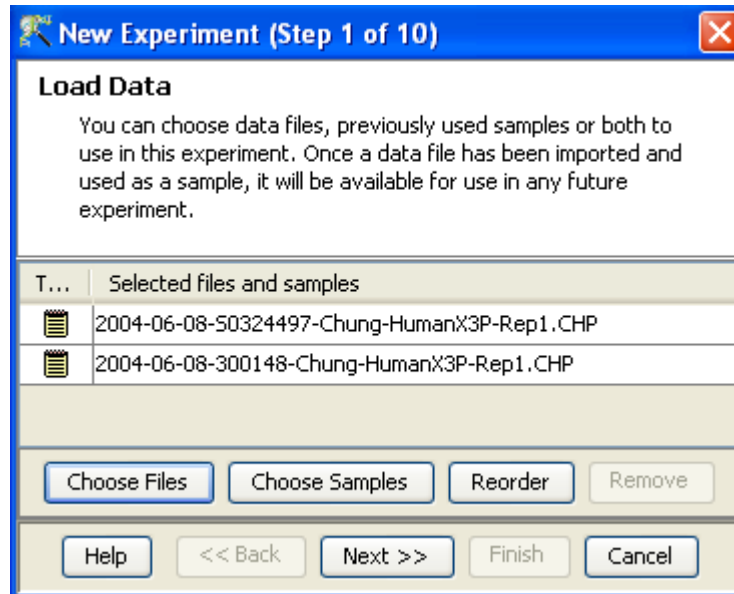


Figure 6.20: Load Data

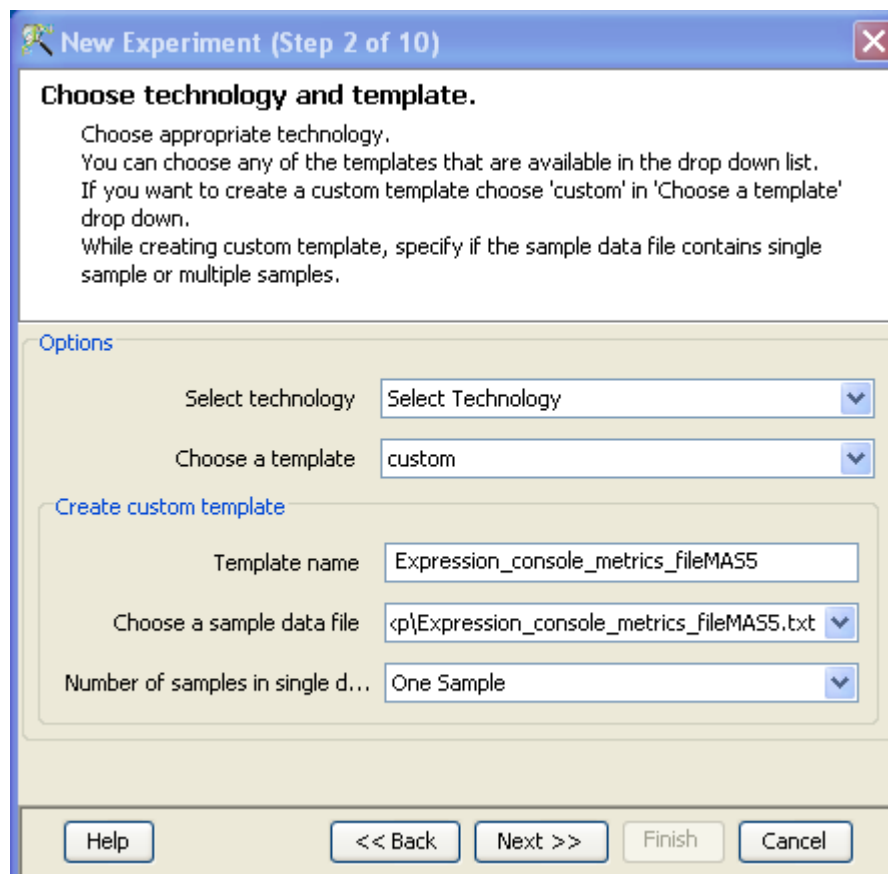


Figure 6.21: Choose Technology and Template

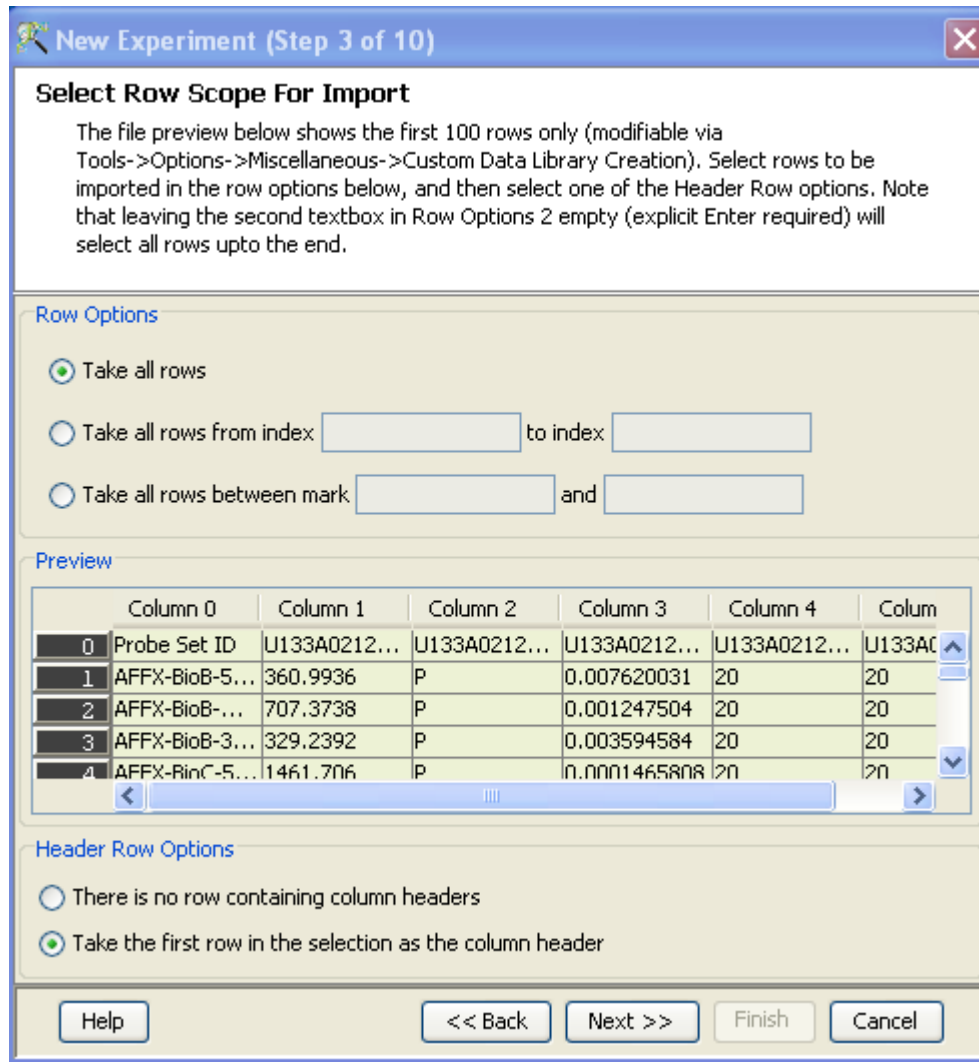


Figure 6.22: Select Row Scope for Import

The **Select Row Scope for Import** window is shown in Figure 6.22.

Step 4 of 10 : Choose Identifier and Signal Column This window allows the user to define the Identifier column, the background signal column and the Flag column from the chosen template file. The flags can be configured.

This step is shown only if the chosen template file has only one sample in the file; for multiple samples in single file, step 5 is shown.

The **Choose Identifier and Signal Column** window is shown in Figure 6.22.

Step 5 of 10 : Single Colour Many Samples in one File Selection This is the equivalent of Step 4 for files with multiple samples. This window provides drop downs to choose Identifier column, the Signal and Flag columns.

There is an option to identify the signal and flag columns using keywords or the user can choose any column and mark it explicitly as signal or flag column. If 'keyword' option is chosen, the user has to

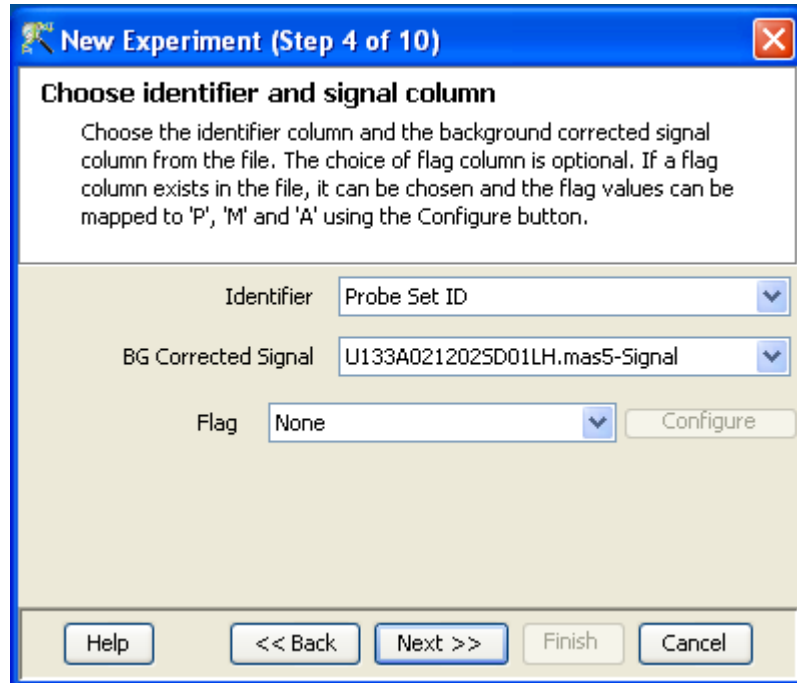


Figure 6.23: Choose Identifier and Signal Column

'Refresh' so that the columns with the specified key words get picked up and shown in the table at the bottom as signal and flag columns.

User can also choose a representative flag column and configure it.

The window is shown in Figure 6.24.

Step 6 of 10 : Select ARR files ARR files are Affymetrix files that hold annotation information for each sample CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample. Click on *Next* to proceed to the next step.

Note that this step is skipped for TXT files.

The **Select ARR files** window is depicted in the Figure 6.25.

Step 7 of 10 : Preprocess Baseline options This step is specific for CEL files. Any one of the Summarization algorithms provided from the drop down menu can be chosen to summarize the data. The available summarization algorithms are:

- The RMA algorithm due to Irizarry et al. [Ir1, Ir2, Bo].
- The MAS5 algorithm, provided by Affymetrix [Hu1].
- The PLIER algorithm due to Hubbell [Hu2].
- The LiWong (dChip) algorithm due to Li and Wong [LiW].
- The GCRMA algorithm due to Wu et al. [Wu].

See Chapter [Probe Summarization Algorithms](#) for details on the above algorithms.

Subsequent to probeset summarization, baseline transformation of the data can be performed. The baseline options include

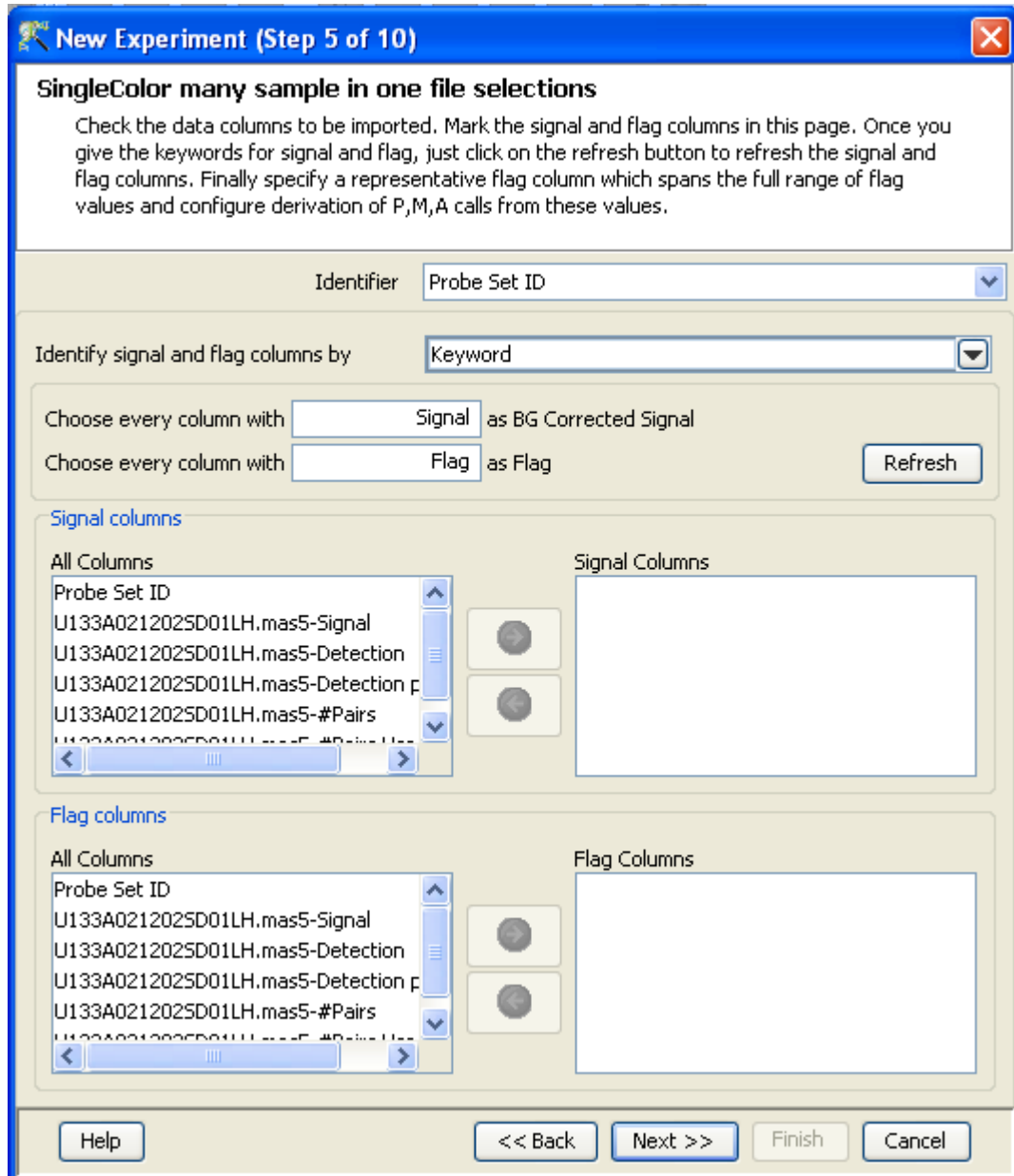


Figure 6.24: Single Colour Many Samples in one File Selection

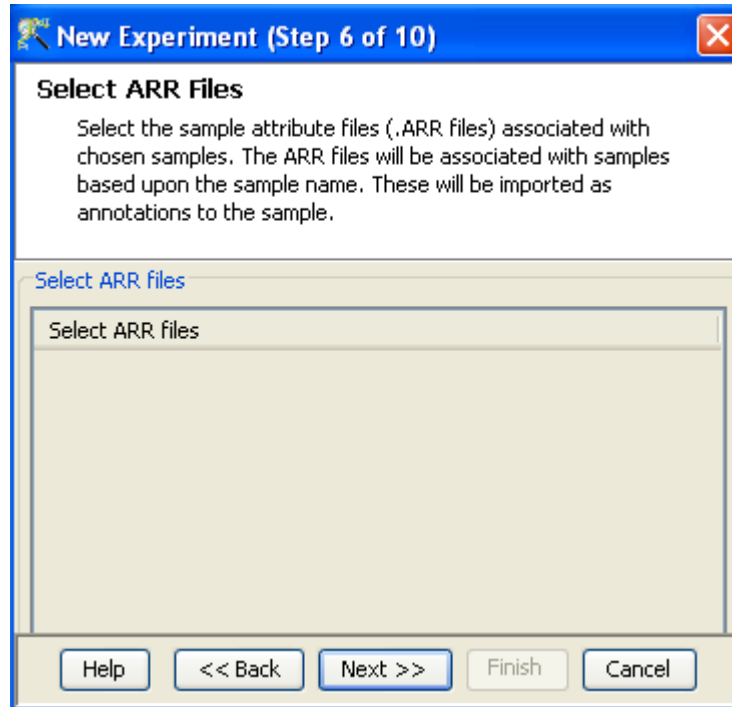


Figure 6.25: Select ARR files

- *Do not perform baseline*
- **Baseline to median of all samples:** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- **Baseline to median of control samples:** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on **Assign value**. This opens up the **Choose Control Samples** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on **OK**. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Figure 6.26 shows the Step to perform base line operations for CEL file in Experiment Creation.

Step 8 of 10 : Normalization This step is specific for CHP files only. See figure ??.

It gives the user the following normalization options.

- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples.

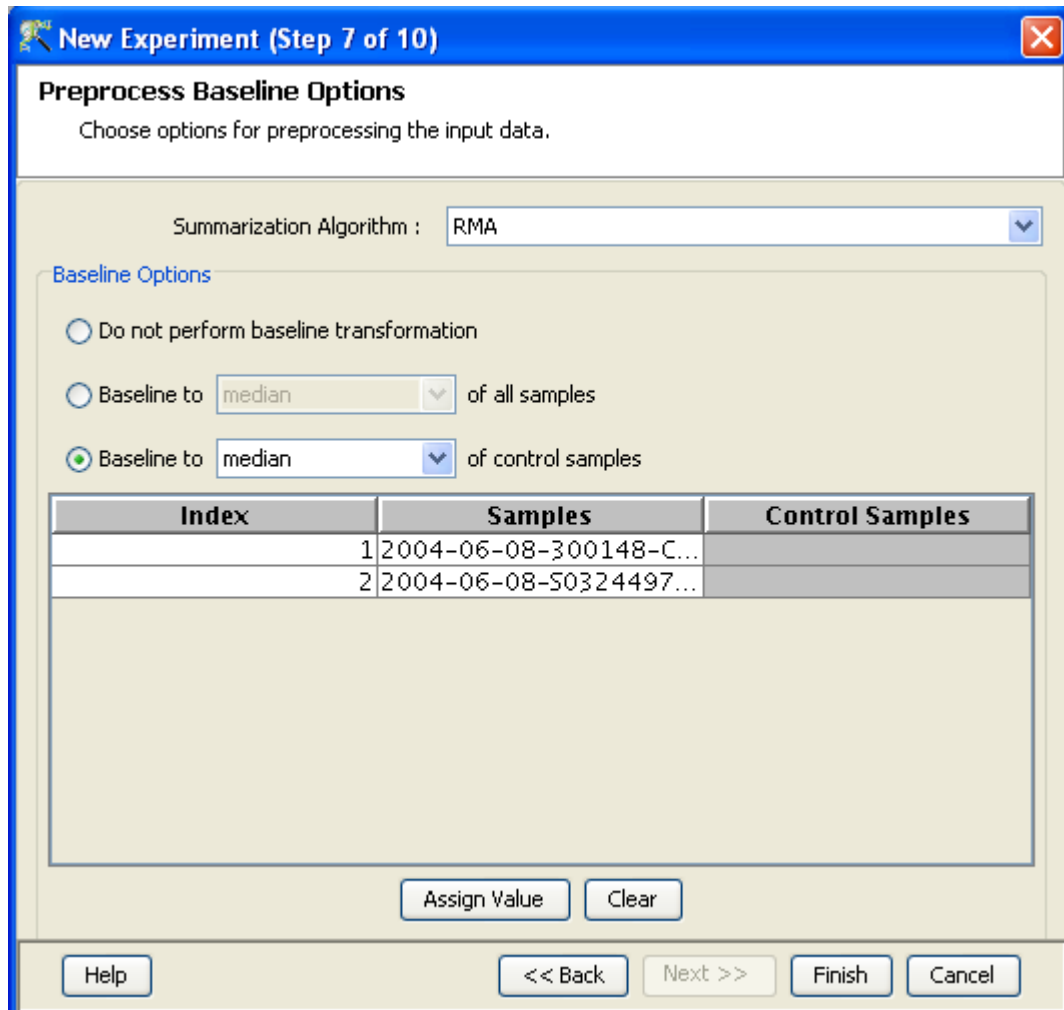


Figure 6.26: Summarization Algorithm

On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the 'Assign Value' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 9 of 10 : Choose entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by

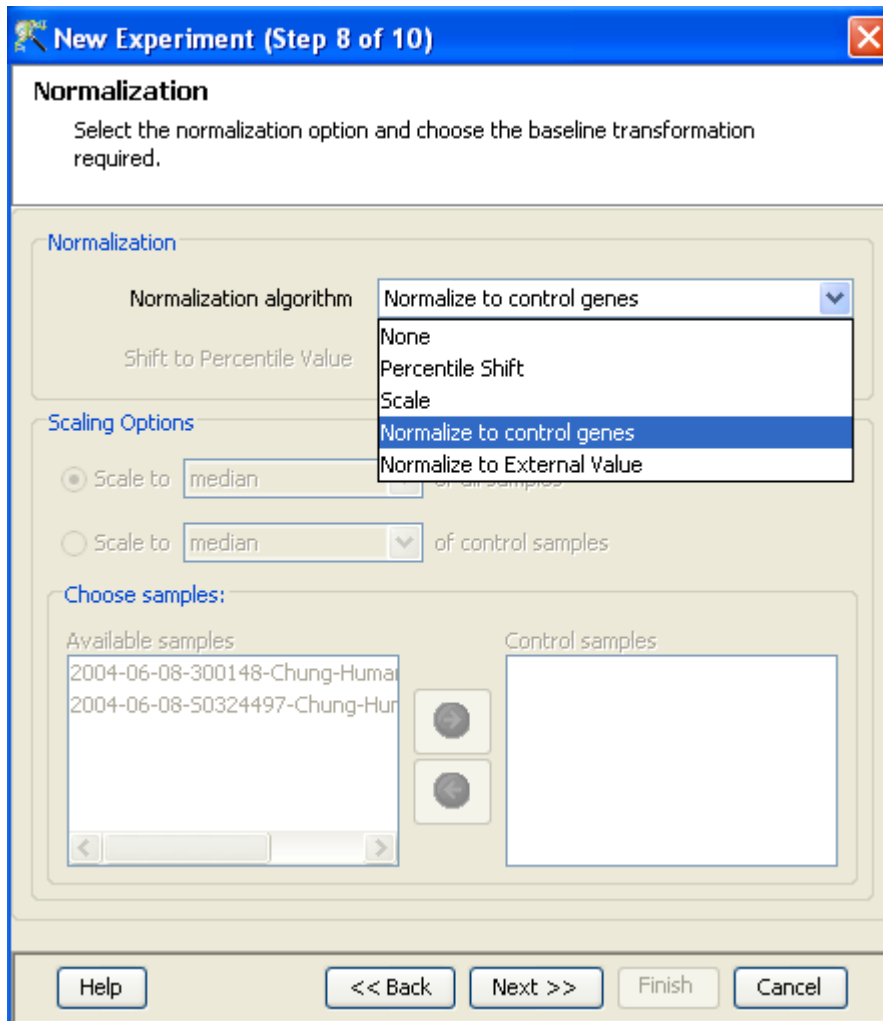


Figure 6.27: Normalization and Baseline Transformation

their probe id. Any other annotation will not be suitable.

- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click **Finish**. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same. See figure 6.28

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

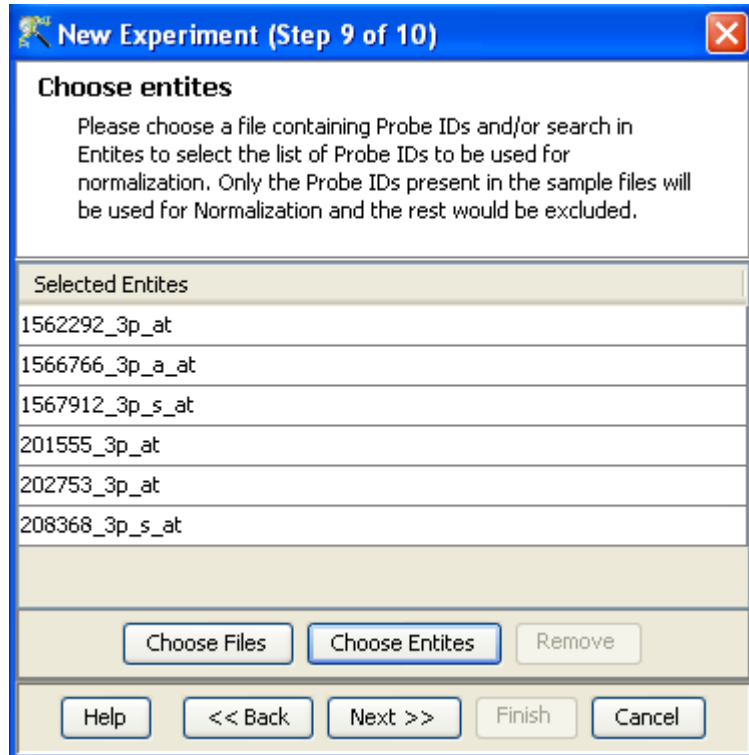


Figure 6.28: Normalize to control genes

Step 10 of 10 : Preprocess Baseline Options (for CHP files) This step allows the user to perform baseline transformation, with CHP files, after normalization. See figure 6.29 The methods available are the same as those used for CEL files in Step 7 of 10.

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to **View** in Toolbar.

Once an experiment is created, the *Advanced Workflow* steps appear on the right hand side. Following is an explanation of the various workflow links:

6.4.2 Experiment Setup

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX** , the *Advanced Workflow*, the method of analysis, the details of the algorithms used and the interpretation of results.
- **Experiment Grouping:** Experiment parameters defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples should be grouped into experimental conditions both for visualization purposes and for analysis. For details refer to the section on [Create Interpretation](#)

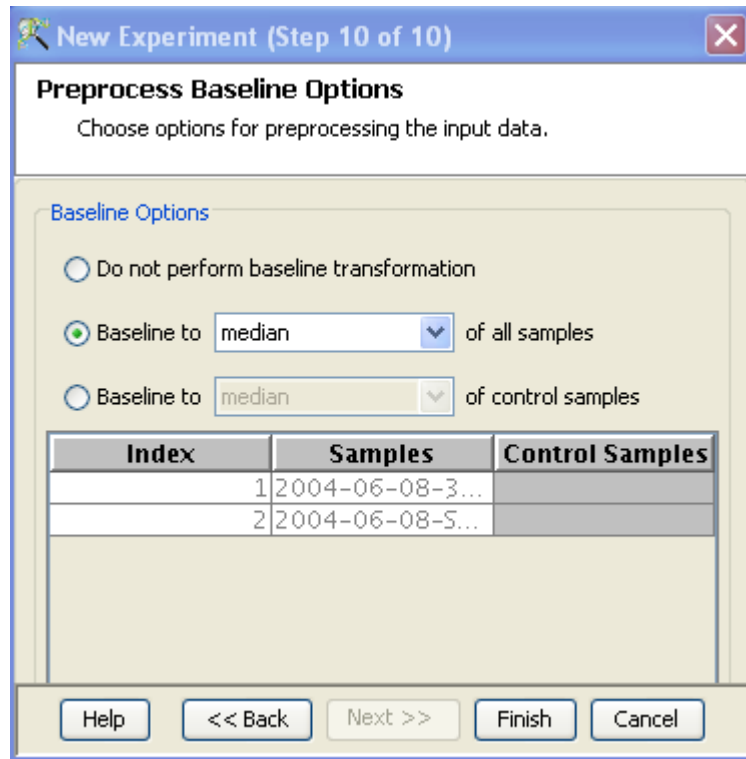


Figure 6.29: Baseline Transformation

- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband

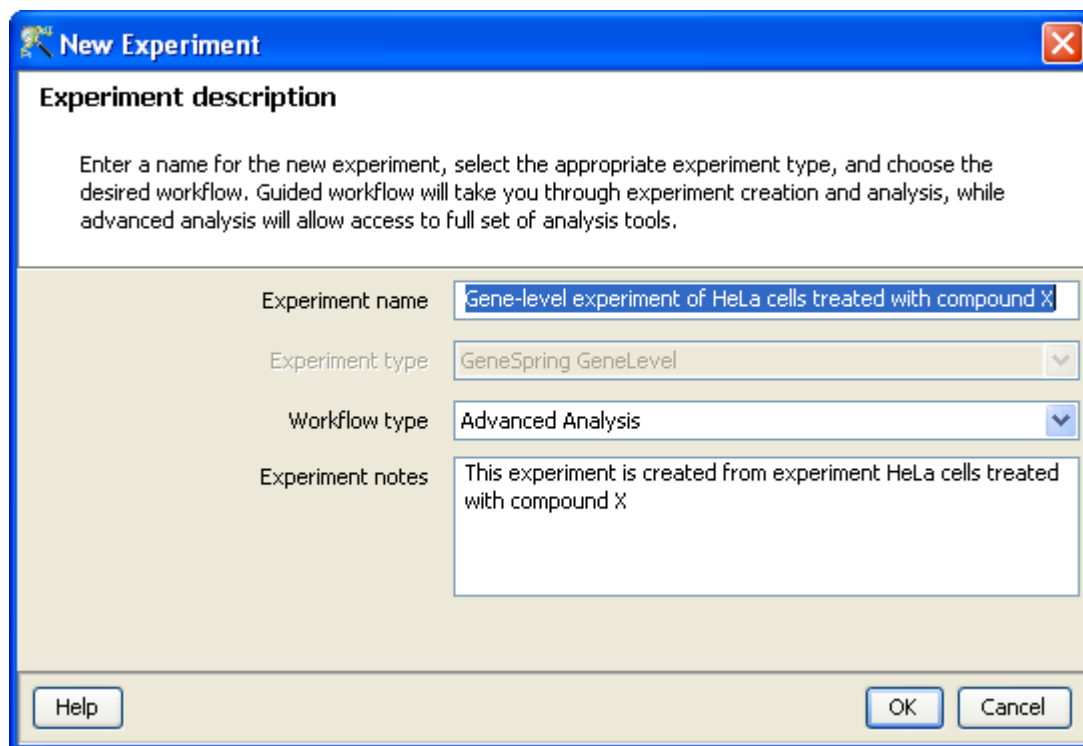


Figure 6.30: Gene Level Experiment Creation

* Probe Sequence

- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of

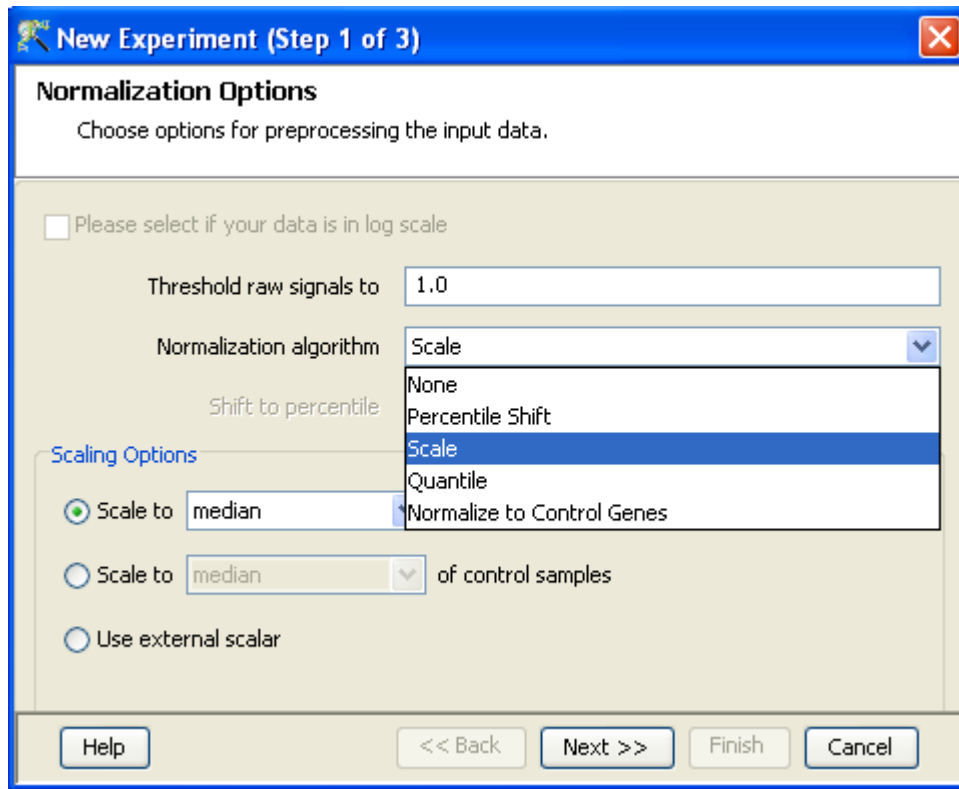


Figure 6.31: Gene Level Experiment Creation - Normalization Options

control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click **Finish**. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

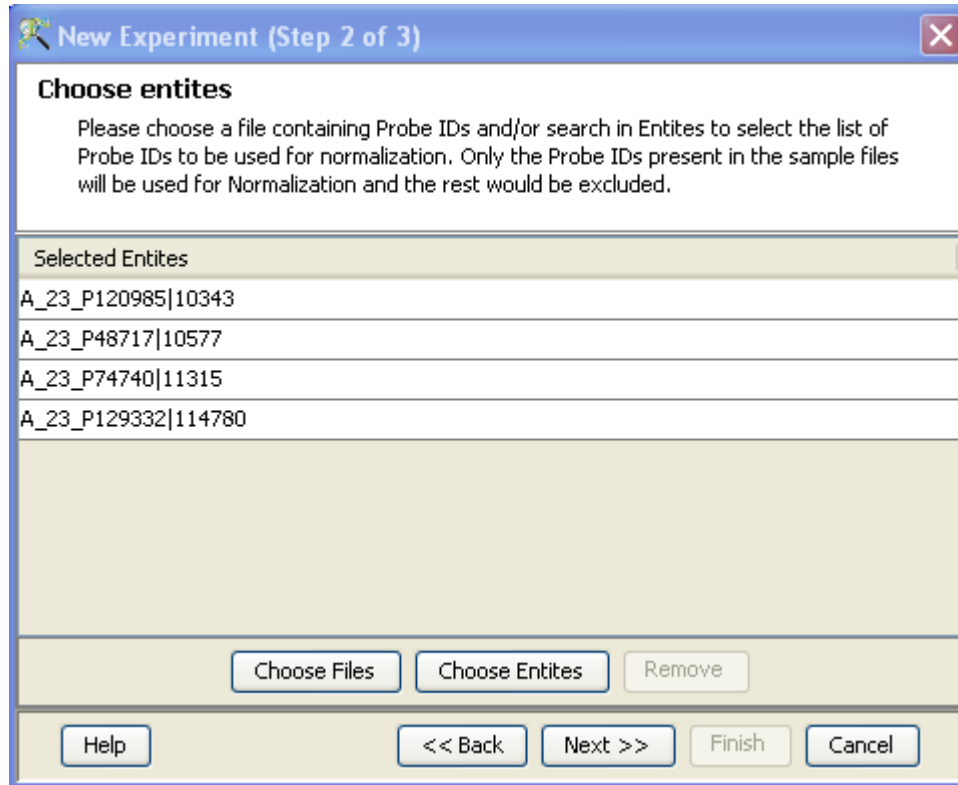


Figure 6.32: Gene Level Experiment Creation - Choose Entities

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

6.4.3 Quality Control

- **Quality Control on Samples:**

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis.

Note that Quality Control is not supported for sample files in TXT format.

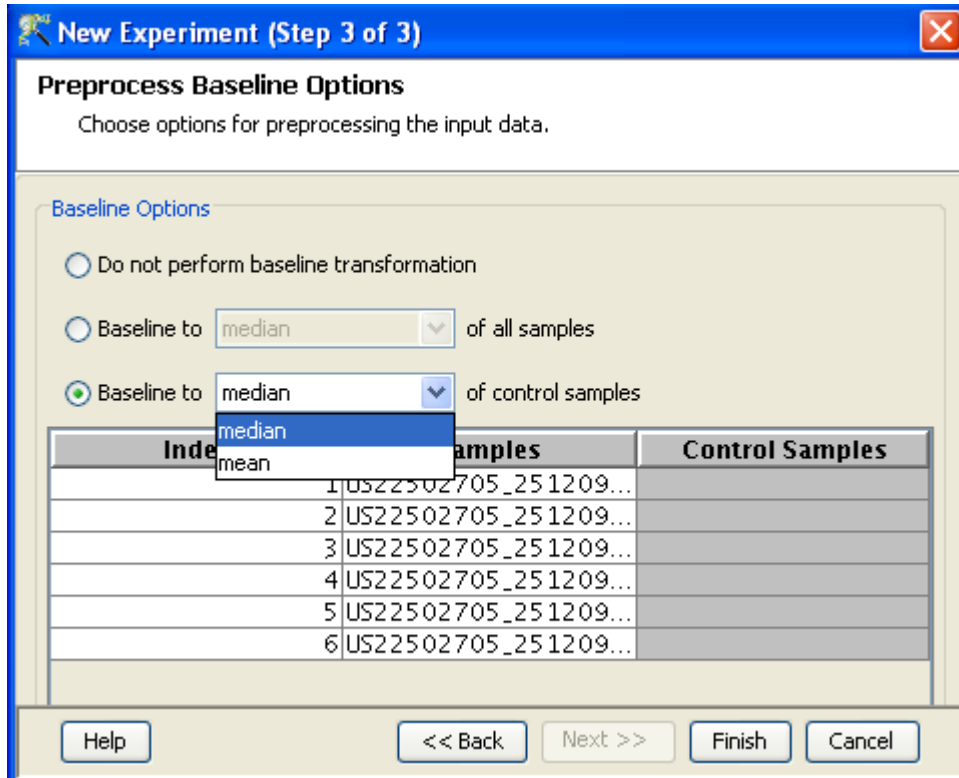


Figure 6.33: Gene Level Experiment Creation - Preprocess Baseline Options

The QC view shows three tiled windows:

- 3D PCA scores, Correlation plots and Correlation Coefficients tabs
- Internal Controls, Hybridization and Experiment grouping
- Legend

Figure 6.34 has the 4 tiled windows which reflect the QC on samples.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. Correlation coefficients are calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value

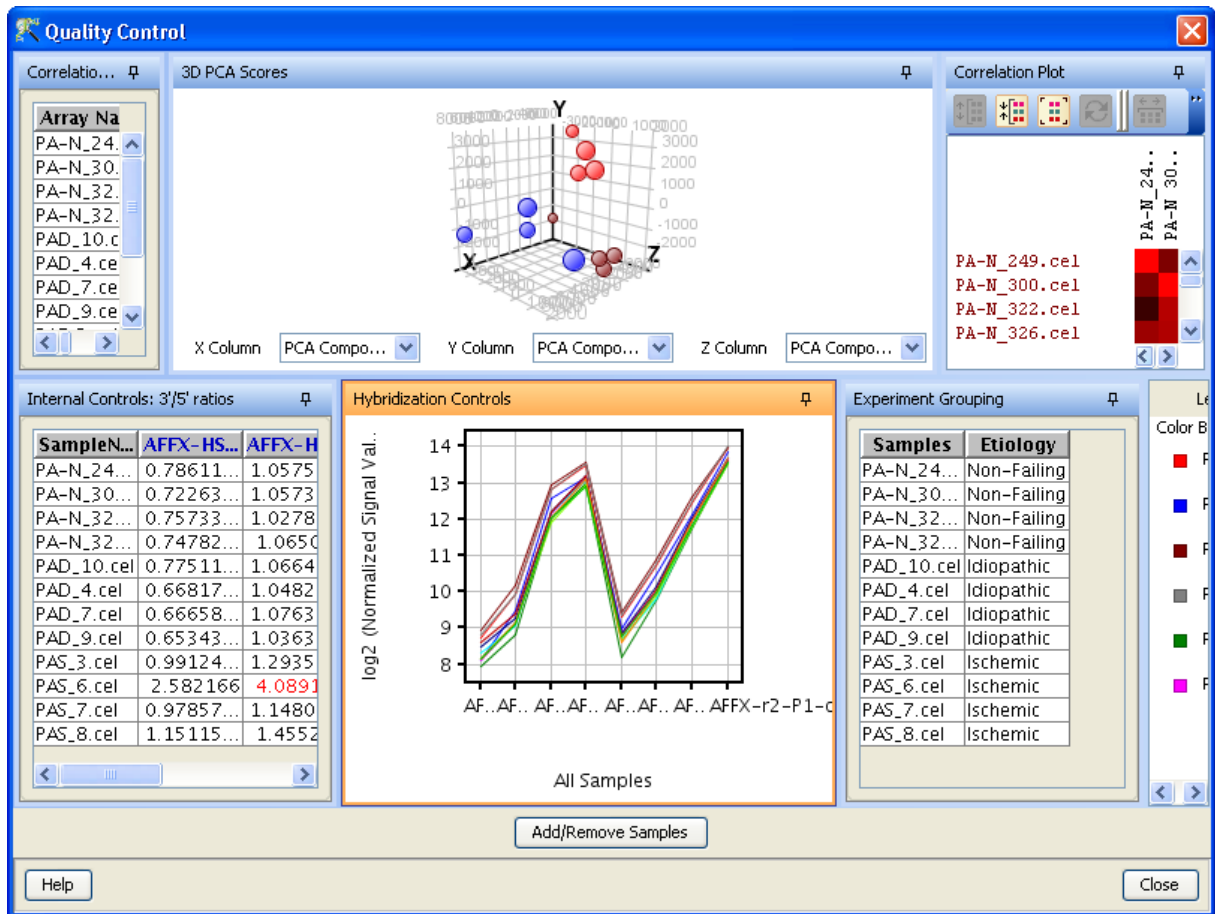


Figure 6.34: Quality Control

from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**.
 Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$

The heatmap is colorable by Experiment Factor information via Right-Click → Properties. Similarly, the intensity levels in the heatmap are also customizable.

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

The *Internal Controls* view depicts RNA sample quality by showing 3'/5' ratios for a set of specific probesets which include the actin and GAPDH probesets. The 3'/5' ratio is output for each such probeset and for each array. The ratios for actin and GAPDH should be no more than 3 (though for *Drosophila*, it should be less than 5). A ratio of more than 3 indicates sample degradation and is indicated in the table in red color.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100 pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be present at least 50%

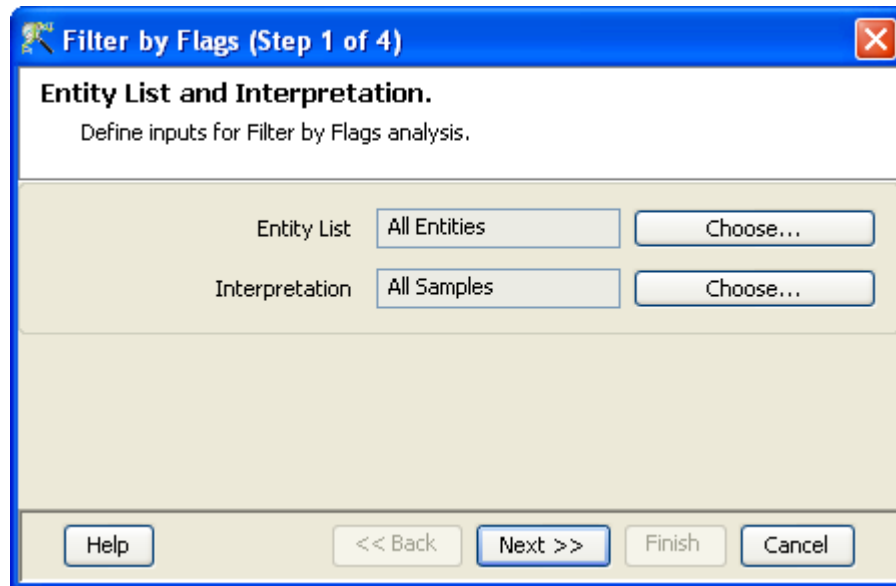


Figure 6.35: Entity list and Interpretation

of the time. *bioC*, *bioD* and *cre* must be Present all of the time and must appear in increasing concentrations. The *Hybridization Controls* shows the signal value profiles of these transcripts (only 3' probesets are taken) where the X axis represents the Biotin labelled cRNA transcripts and the Y axis represents the log of the Normalized Signal Values.

Experiment Grouping tab shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-summation of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on *OK* to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags:**

This step is specific for analysis where MAS5.0 summarization has been done on samples. MAS5.0 generates flag values, the P(present), M(marginal) and A(absent), for each row in each sample. In the *Filter Probe Set by Flags* step, entities can be filtered based on their flag values. This is done in 4 steps:

1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.
2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.

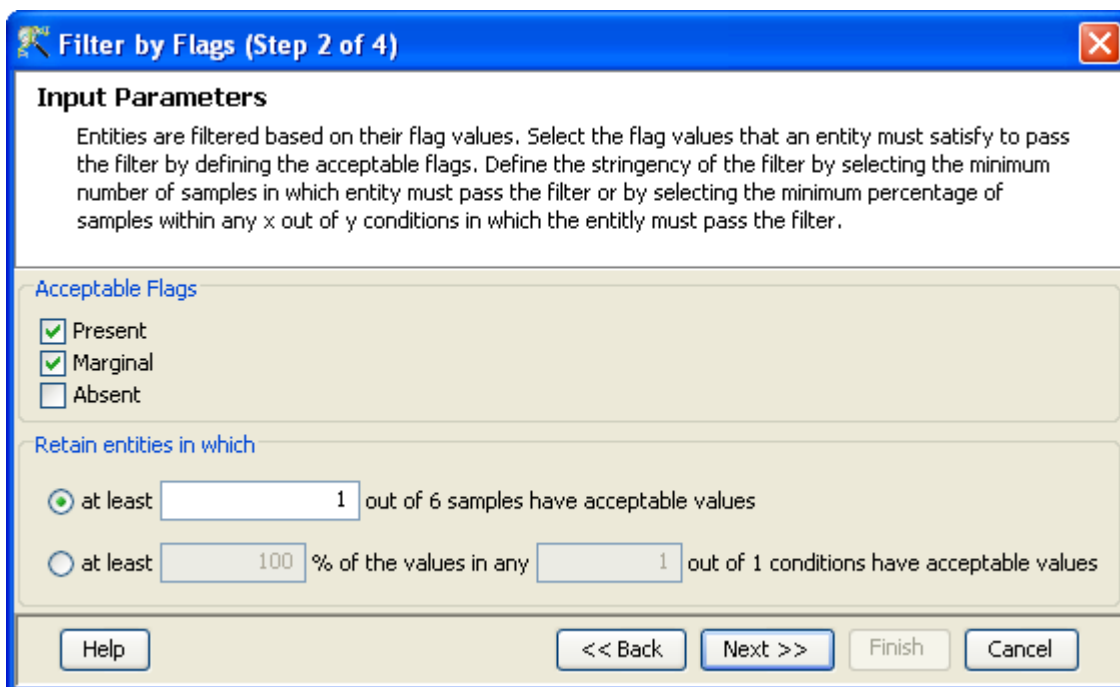


Figure 6.36: Input Parameters

3. Step 3 of 4: A spreadsheet and a profile plot appear as two tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. (See Figure 6.37).
 4. Step 4 of 4: Click *Next* to annotate and save the entity list. (See Figure 6.38).
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

6.4.4 Analysis

- **Statistical Analysis**
For details refer to section [Statistical Analysis](#) in the advanced workflow.
- **Filter on Volcano Plot**
For details refer to section [Filter on Volcano Plot](#)
- **Fold Change**
For details refer to section [Fold Change](#)
- **Clustering**
For details refer to section [Clustering](#)

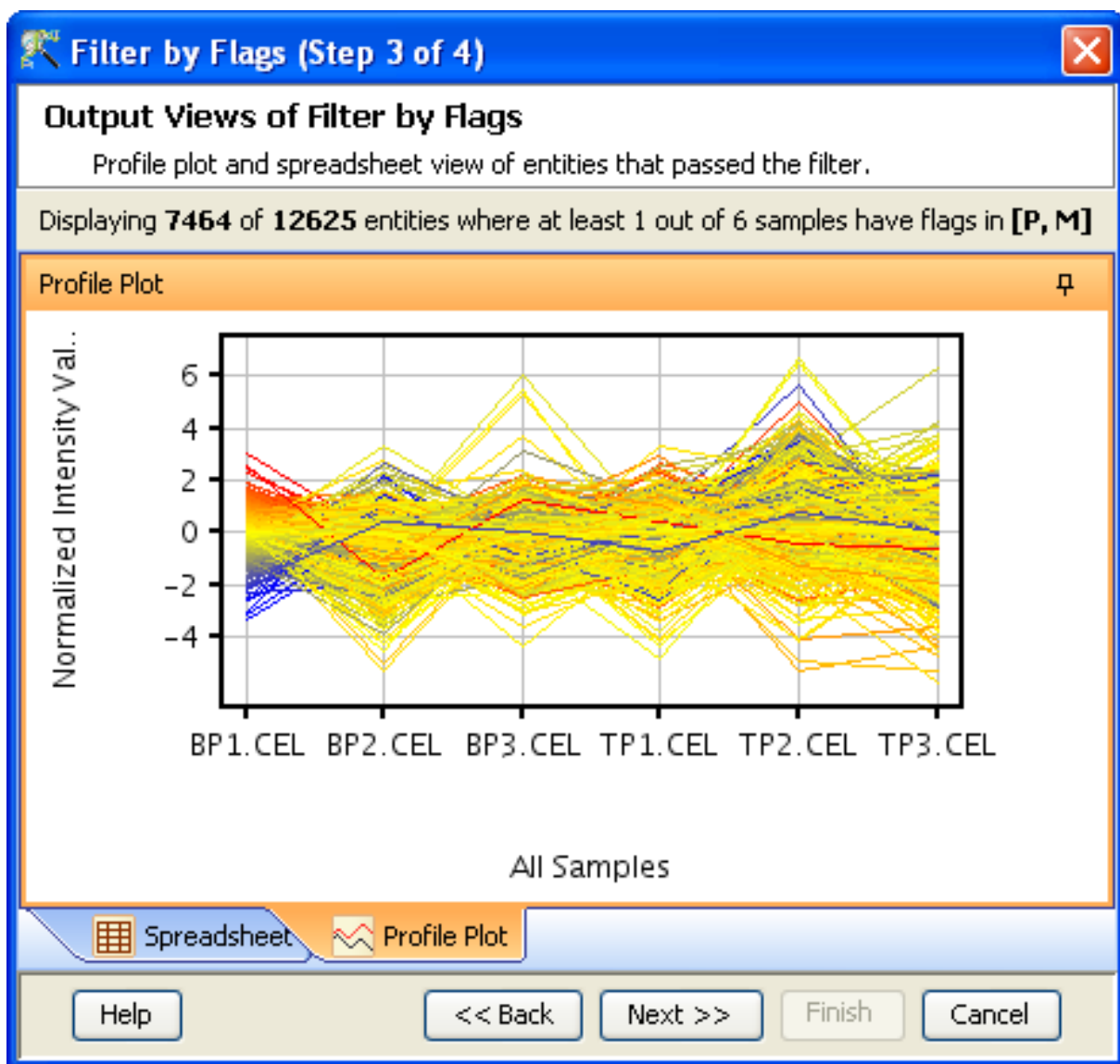


Figure 6.37: Output Views of Filter by Flags

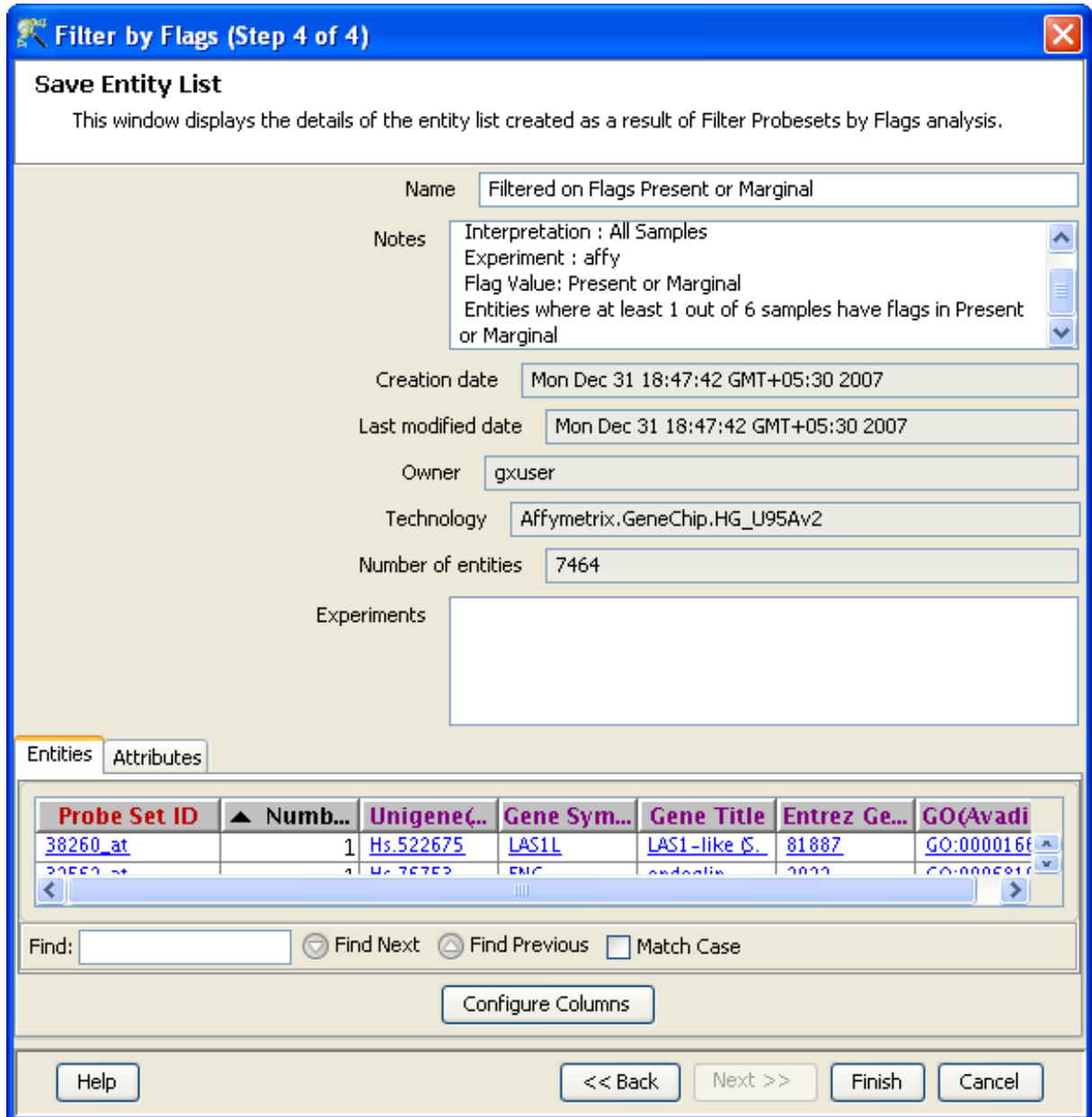


Figure 6.38: Save Entity List

- **Find Similar Entities**
For details refer to section [Find Similar Entities](#)
- **Filter on Parameters**
For details refer to section [Filter on Parameters](#)
- **Principal Component Analysis**
For details refer to section [PCA](#)

6.4.5 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

6.4.6 Results

- **Gene Ontology (GO) analysis**
GO is discussed in a separate chapter called [Gene Ontology Analysis](#).
- **Gene Set Enrichment Analysis (GSEA)**
Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).
- **Gene Set Analysis (GSA)**
Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).
- **Pathway Analysis**
Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).
- **Find Similar Entity Lists**
This feature is discussed in a separate section called [Find Similar Entity Lists](#)
- **Find Significant Pathways**
This feature is discussed in a separate section called [Find Significant Pathways](#).
- **Launch IPA**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Import IPA Entity List**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Extract Interactions via NLP**
This feature is discussed in detail in the chapter [Pathway Analysis](#).

6.4.7 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)
- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)
- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)
- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

6.4.8 Affymetrix Technology creation using Custom CDF

Creating a Technology using Affymetrix Custom CDF:

GeneSpring GX offers the user a facility to create Custom Affymetrix expression (GeneChip) technology if you have a Custom CDF file. This happens in situations where you have a Custom Affymetrix array or might want to use a Custom CDF for a Standard technology, for e.g., the ones obtained from http://brainarray.mbni.med.umich.edu/brainarray/Database/CustomCDF/genomic_curated_CDF.asp

Following are the steps for creating an Affymetrix Custom technology:

1. Go to *Annotations* → *Create Technology* → *Affymetrix Expression*.
2. For creating a Custom Affymetrix technology, the CDF file is mandatory. The PSI, CIF, Probe Tab and annotation files are optional. If however, Probe Tab is also being used, then make sure that the system has 'R' package installed and that its path has been set, by going to *Tools* → *Options* → *Miscellaneous* → *R path*. The Bioconductor packages `makecdfenv`, `matchprobes` and `gcrma` also need to be installed before the Probe Tab file can be used. See Figures [6.39](#) and [6.40](#).
3. The CDF file name should reflect the GeneChip name for which it is being used. If the Custom CDF is derived from a Standard technology, then it should be renamed to that of the Standard technology for e.g., the Custom CDF file `HS95Av2_HS_UG_1.cdf` derived from GeneChip `HG_U95Av2` should be renamed to `HG_U95Av2.cdf`. This is necessary because to ensure that no errors occur, **GeneSpring GX** tries to match the CDF/technology name with the GeneChip name from the data file, during the process of experiment creation. Taking again the example of `HS95Av2_HS_UG_1.cdf` (for the GeneChip `HG_U95Av2`), if the CDF is not renamed, an experiment created using the `HG_U95Av2` CEL files will use the the Standard `Affymetrix.GeneChip.HG_U95Av2` technology, instead of the newly created `Affymetrix.GeneChip.HS95Av2_HS_UG_1` technology.
4. In case of Custom CDF derived from a Standard technology, refer to either *Search* → *Technologies* or go to *Annotations* → *Create Technology* → *From Agilent Server* to get the exact name of the GeneChip (It is case sensitive).

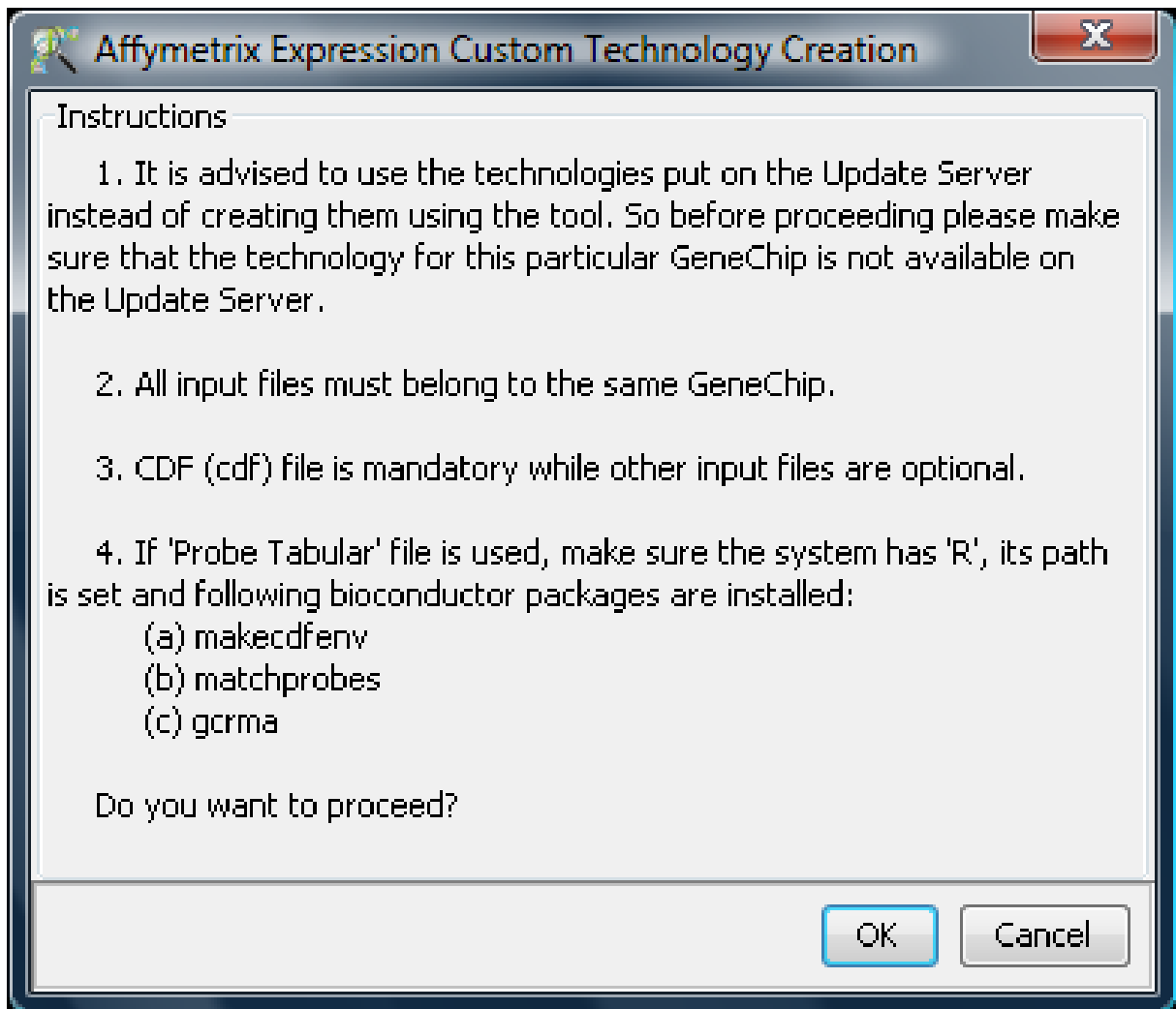


Figure 6.39: Confirmation Dialog Box

5. The technology created will automatically take the name of the Custom CDF and will be named as `Affymetrix.GeneChip.¡CDF_file_name¡`. If a technology with the same name, standard or custom already exists, then **GeneSpring GX** overwrites it with the new one after the user's confirmation.

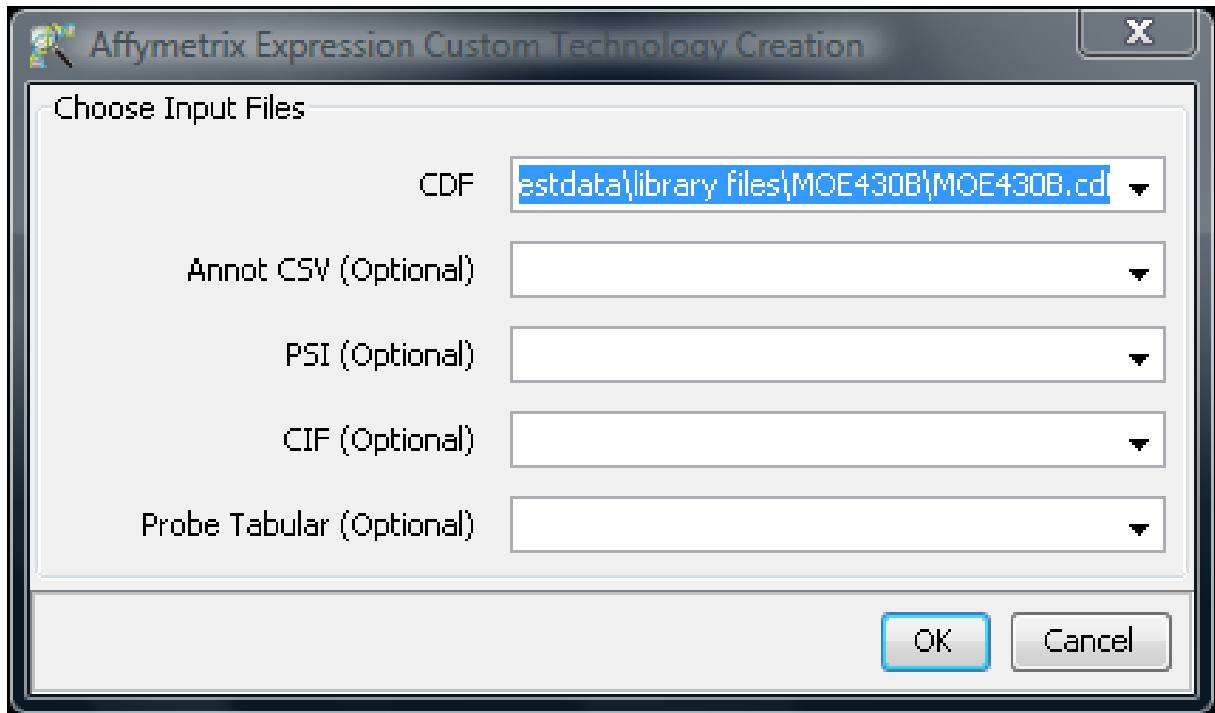


Figure 6.40: Choose Input Files

Notes:

1. In **GeneSpring GX**, for a given Affymetrix GeneChip, at any point of time, you cannot have more than one technology.
2. If you replace an older Affymetrix technology with the new one, then the behavior of the experiments created with the older technology is unpredictable. So it is advised to remove/delete those.
3. If an experiment needs to be analyzed now with the standard technology for which a Custom CDF had been used, it will be necessary to delete the technology created using the Custom CDF. Otherwise, the standard technology cannot be downloaded.

Chapter 7

Affymetrix Summarization Algorithms

This section describes technical details of the various probe summarization algorithms, normalization using spike-in and housekeeping probesets, and computing absolute calls.

7.0.1 Probe Summarization Algorithms

Probe summarization algorithms perform the following 3 key tasks: Background Correction, Normalization, and Probe Summarization (i.e. conversion of probe level values to probeset expression values in a robust, i.e., outlier resistant manner. The order of the last two steps could differ for different probe summarization algorithms. For example, the RMA algorithm does normalization first, while MAS5 does normalization last. In RMA and GCRMA the summarization is inherently on log scale, whereas in PLIER and MAS5 summarization works on linear scale. Further, the methods mentioned below fall into one of two classes – the PM based methods and the $PM - MM$ based methods. The $PM - MM$ based methods take $PM - MM$ as their measure of background corrected expression while the PM based measures use other techniques for background correction. MAS5, MAS4, and Li-Wong are $PM - MM$ based measures while RMA and GCRMA are PM based measures. For a comparative analysis of these methods, see [33, 34] or [1].

A brief description of each of the probe summarization options available in **GeneSpring GX** is given below. Some of these algorithms are native implementations within **GeneSpring GX** and some are directly based on the Affymetrix codebase. The exact details are described in the table below.

RMA with only pm probes	Implemented in GeneSpring GX	Validated against R with bgversion=2
GCRMA	Implemented in GeneSpring GX	Validated against default GCRMA in R
MAS5	Licensed from Affymetrix	Validated against Affymetrix Data
PLIER	Summarization licensed from Affymetrix, Normalization implemented in GeneSpring GX	Validated against Affymetrix Data
LiWong	Implemented in GeneSpring GX	Validated against R
Absolute Calls	Licensed from Affymetrix	Validated against Affymetrix Data

Masked Probes and Outliers. Finally, note that CEL files have masking and outlier information about certain probes. These masked probes and outliers are removed.

The RMA (Robust Multichip Averaging) Algorithm

The RMA method was introduced by Irizarry et al. [33, 34] and is used as part of the RMA package in the Bioconductor suite. In contrast to MAS5, this is a PM based method. It has the following components.

Background Correction. The RMA background correction method is based on the distribution of PM values amongst probes on an Affymetrix array. The key observation is that the smoothed histogram of the $\log(PM)$ values exhibits a sharp normal-like distribution to the left of the mode (i.e., the peak value) but stretches out much more to the right, suggesting that the PM values are a mixture of non-specific binding and background noise on one hand and specific binding on the other hand. The above peak value is a natural estimate of the average background noise and this can be subtracted from all PM values to get background corrected PM values. However, this causes the problem of negative values. Irizarry et al. [33, 34] solve the problem of negative values by imposing a positive distribution on the background corrected values. They assume that each observed PM value O is a sum of two components, a signal S which is assumed to be exponentially distributed (and is therefore always positive) and a noise component N which is normally distributed. The background corrected value is obtained by determining the expectation of S conditioned on O which can be computed using a closed form formula. However, this requires estimating the decay parameter of the exponential distribution and the mean and variance of the normal distribution from the data at hand. These are currently estimated in a somewhat ad-hoc manner.

Normalization. The RMA method uses Quantile normalization. Each array contains a certain distribution of expression values and this method aims at making the distributions across various arrays not just similar but identical! This is done as follows. Imagine that the expression values from various arrays have been loaded into a dataset with probesets along rows and arrays along columns. First, each column is sorted in increasing order. Next, the value in each row is replaced with the average of the values in this row. Finally, the columns are unsorted (i.e., the effect of the sorting step is reversed so that the items

in a column go back to wherever they came from). Statistically, this method seems to obtain very sharp normalizations [10]. Further, implementations of this method run very fast.

GeneSpring GX uses all arrays to perform normalization on the raw intensities, irrespective of their variance.

Probe Summarization. RMA models the observed probe behavior (i.e., $\log(PM)$ after background correction) on the log scale as the sum of a probe specific term, the actual expression value on the log scale, and an independent identically distributed noise term. It then estimates the actual expression value from this model using a robust procedure called *Median Polish*, a classic method due to Tukey.

The GCRMA Algorithm

This algorithm was introduced by Wu et al [52] and differs from RMA only in the background correction step. The goal behind its design was to reduce the bias caused by not subtracting MM in the RMA algorithm. The GCRMA algorithm uses a rather technical procedure to reduce this bias and is based on the fact that the non-specific affinity of a probe is related to its base sequence. The algorithm computes a background value to be subtracted from each probe using its base sequence

The Li-Wong Algorithm

There are two versions of the Li-Wong algorithm [38], one which is $PM - MM$ based and the other which is PM based. Both are available in the dChip software. **GeneSpring GX** has only the $PM - MM$ version.

Background Correction. No special background correction is used by the **GeneSpring GX** implementation of this method. Some background correction is implicit in the $PM - MM$ measure.

Normalization. While no specific normalization method is part of the Li-Wong algorithm as such, dChip uses *Invariant Set* normalization. An *invariant set* is a collection of probes with the most conserved ranks of expression values across all arrays. These are identified and then used very much as spike-in probesets would be used for normalization across arrays. In **GeneSpring GX**, the current implementation uses Quantile Normalization [10] instead, as in RMA.

Probe Summarization. The Li and Wong [38] model is similar to the RMA model but on a linear scale. Observed probe behavior (i.e., $PM - MM$ values) is modelled on the linear scale as a *product* of a probe affinity term and an actual expression term along with an additive normally distributed independent error term. The maximum likelihood estimate of the actual expression level is then determined using an estimation procedure which has rules for outlier removal. The outlier removal happens at multiple levels. At the first level, outlier arrays are determined and removed. At the second level, a probe is removed from all the arrays. At the third level, the expression value for a particular probe on a particular array is

rejected. These three levels are performed in various iterative cycles until convergence is achieved. Finally, note that since $PM - MM$ values could be negative and since **GeneSpring GX** outputs values always on the logarithmic scale, negative values are thresholded to 1 before output.

The Average Difference and Tukey-BiWeight Algorithms

These algorithms are similar to the MAS4 and MAS5 methods [30] used in the Affymetrix software, respectively.

Background Correction. These algorithm divide the entire array into 16 rectangular zones and the second percentile of the probe values in each zone (both PM's and MM's combined) is chosen as the background value for that region. For each probe, the intention now is to reduce the expression level measured for this probe by an amount equal to the background level computed for the zone containing this probe. However, this could result in discontinuities at zone boundaries. To make these transitions smooth, what is actually subtracted from each probe is a weighted combination of the background levels computed above for all the zones. Negative values are avoided by thresholding.

Probe Summarization. The one-step Tukey Biweight algorithm combines together the background corrected $\log(PM - MM)$ values for probes within a probe set (actually, a slight variant of MM is used to ensure that $PM - MM$ does not become negative). This method involves finding the median and weighting the items based on their distance from the median so that items further away from the median are down-weighted prior to averaging.

The Average Difference algorithm works on the background corrected $PM - MM$ values for a probe. It ignores probes with $PM - MM$ intensities in the extreme 10 percentiles. It then computes the mean and standard deviation of the $PM - MM$ for the remaining probes. Average of $PM - MM$ intensities within 2 standard deviations from the computed mean is thresholded to 1 and converted to the log scale. This value is then output for the probeset.

Normalization. This step is done after probe summarization and is just a simple scaling to equalize means or trimmed means (means calculated after removing very low and very high intensities for robustness).

The PLIER Algorithm

This algorithm was introduced by Hubbell [31] and introduces a integrated and mathematically elegant paradigm for background correction and probe summarization. The normalization performed is the same as in RMA, i.e., Quantile Normalization. After normalization, the PLIER procedure runs an optimization procedure which determines the best set of weights on the PM and MM for each probe pair. The goal is to weight the PMs and MMs differentially so that the weighted difference between PM and MM is non-negative. Optimization is required to make sure that the weights are as close to 1 as possible. In the process of determining these weights, the method also computes the final summarized value.

Comparative Performance

For comparative performances of the above mentioned algorithm, see [33, 34] where it is reported that the RMA algorithm outperforms the others on the GeneLogic spike-in study [26]. Alternatively, see [1] where all algorithms are evaluated against a variety of performance criteria.

7.0.2 Computing Absolute Calls

GeneSpring GX uses code licenced from Affymetrix to compute calls. The Present, Absent and Marginal Absolute calls are computed using a Wilcoxon Signed Rank test on the $(PM-MM)/(PM+MM)$ values for probes within a probeset. This algorithm uses the following parameters for making these calls:

- The *Threshold Discrimination Score* is used in the Wilcoxon Signed Rank test performed on $(PM-MM)/(PM+MM)$ values to determine signs. A higher threshold would decrease the number of false positives but would increase the number of false negatives.
- The second and third parameters are the *Lower Critical p-value* and the *Higher Critical p-value* for making the calls. Genes with p-value in between these two values will be called Marginal, genes with p-value above the Higher Critical p-value will be called Absent and all other genes will be called Present.

Parameters for Summarization Algorithms and Calls

The algorithms MAS5 and PLIER and the Absolute Call generation procedure use parameters which can be seen at *File* \rightarrow *Configuration*. However, modifications of these parameters are not currently available in **GeneSpring GX**. These should be available in the future versions.

Chapter 8

Analyzing Affymetrix Exon Expression Data

Affymetrix Exon chips are being increasingly used for assessing the expression levels of transcripts. **GeneSpring GX** supports this Affymetrix Exon Expression Technology.

8.1 Running the Affymetrix Exon Workflow

Upon launching **GeneSpring GX** , the startup is displayed with 3 options.

- **Create new project**
- **Open existing project**
- **Open recent project**

Either a new project can be created or a previously generated project can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on **OK** to proceed.

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment:** This allows the user to create a new experiment. (steps described below).
2. **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

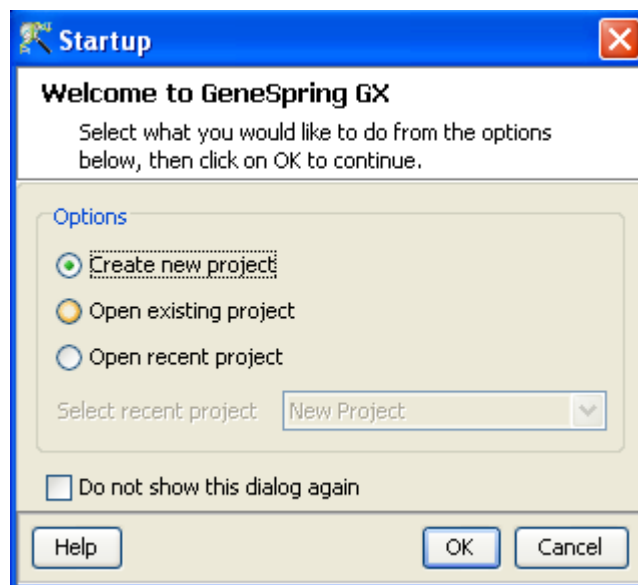


Figure 8.1: Welcome Screen

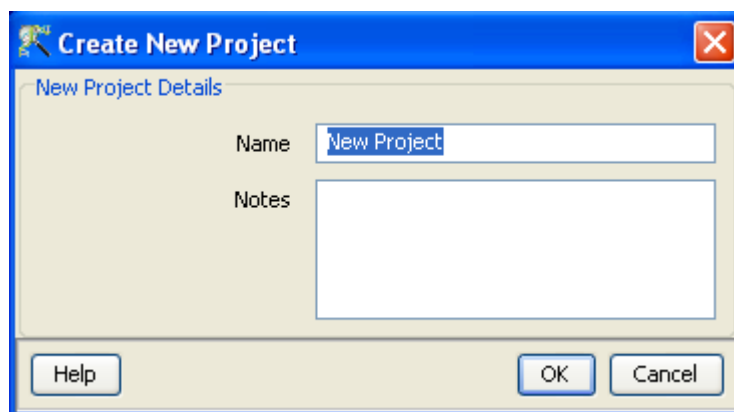


Figure 8.2: Create New project

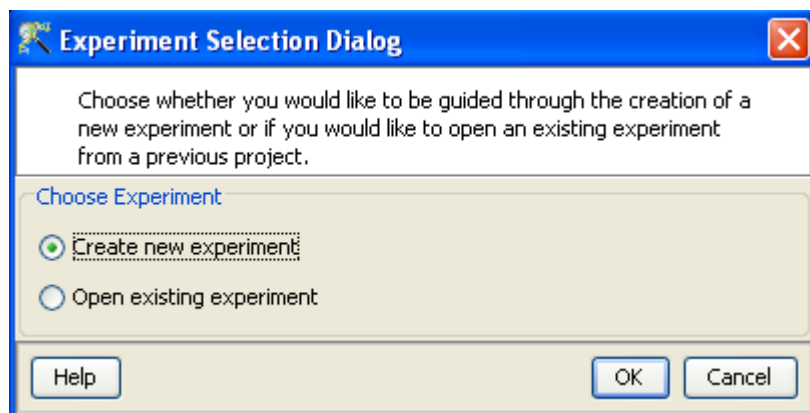


Figure 8.3: Experiment Selection

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the experiment type gives the user the option to choose between the multiple experiment types namely Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Agilent miRNA, Generic Single Color, Generic Two Color, Pathway and RealTime-PCR experiment.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Guided Workflow**
2. **Advanced Analysis Workflow**

Guided Workflow is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Advanced Analysis**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples
3. Reorder
4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from

a scanner. On the other hand, a sample is created within **GeneSpring GX**, when it associates the data files with its appropriate technology (See the section on [Technology](#)). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the *Choose Samples* option. For selecting data files and creating an experiment, click on the *Choose File(s)* button, navigate to the appropriate folder and select the files of interest. Click on *OK* to proceed.

The technology specific for any chip type needs to be created or downloaded only once. Thus, upon creating an experiment of a specific chip type for the first time, **GeneSpring GX** prompts the user to download the technology from the update server. If an experiment has been created previously with the same technology, **GeneSpring GX** then directly proceeds with experiment creation. Clicking on the *Choose Samples* button, opens a sample search wizard, with the following search conditions:

1. **Search field:** Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition:** Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on *Add* button, or can be removed from the list using *Remove* button.

Figures [8.4](#), [8.5](#), [8.6](#), [8.7](#) show the process of choosing experiment type, loading data, choosing samples and re-ordering the data files.

8.2 Data Processing

1. **File formats:** The data file should be present either as a CEL file or a CHP file. However while creating an experiment; only one type of file (CEL/CHP) can be used.
2. **Raw signal values (CEL files):** In an Affymetrix Exon Expression experiment, the term "raw" signal values refers to the linear data which has been summarized using a summarization algorithm (RMA16, PLIER 16 and Iterative PLIER 16). All summarization algorithms also do variance stabilization by adding 16.
3. **Raw signal values (CHP files):** In an Affymetrix Exon Expression experiment, the term "raw" files refers to the linear data obtained from the CHP files.
4. **Normalized signal values (CEL files):** "Normalized" values are generated after the log transformation and baseline transformation step.

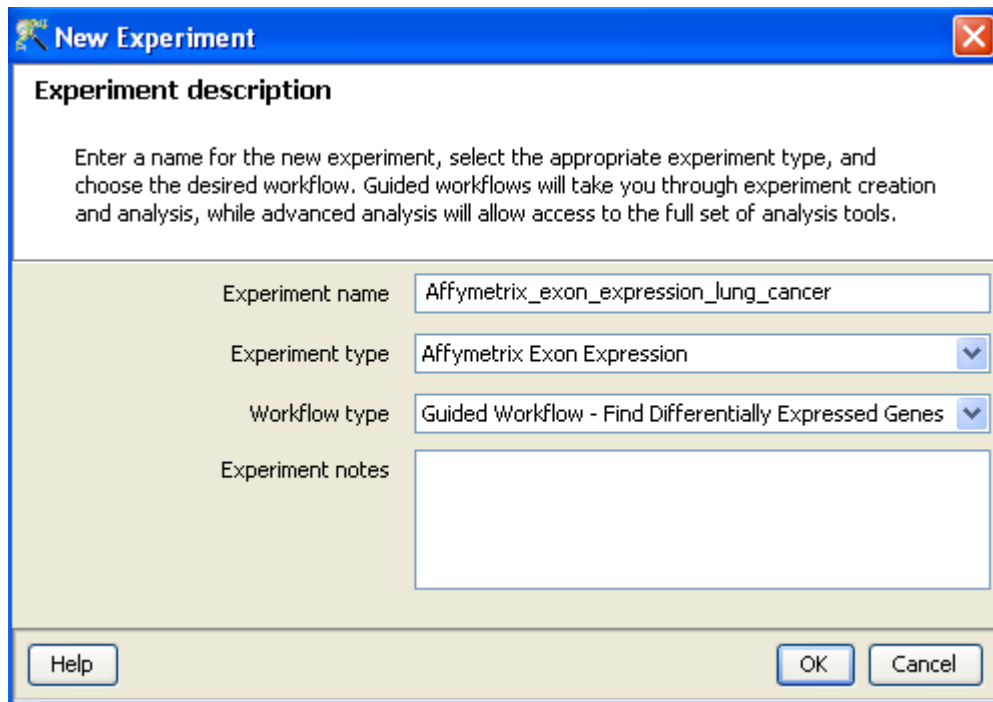


Figure 8.4: Experiment Description

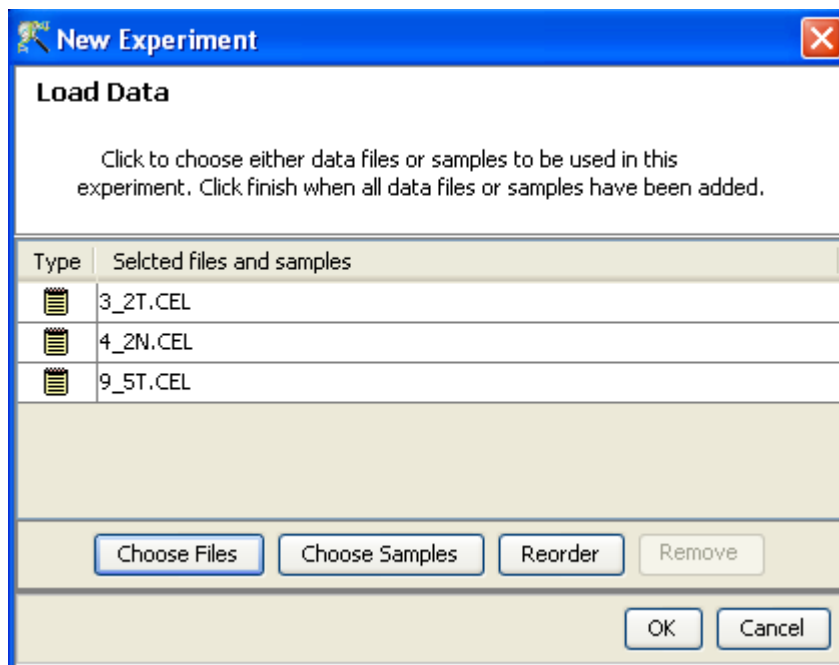


Figure 8.5: Load Data

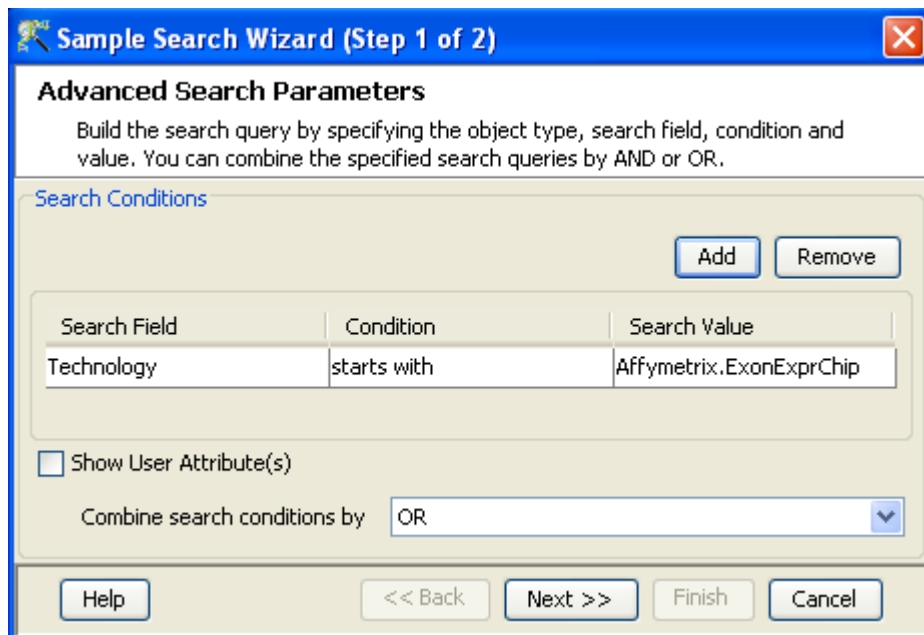


Figure 8.6: Choose Samples

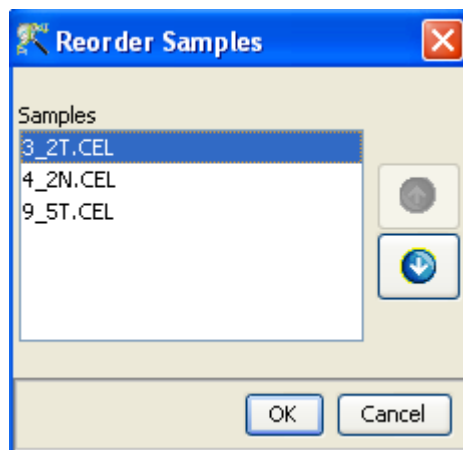


Figure 8.7: Reordering Samples

5. **Normalized signal values (CHP files):** The term "Normalized" refers to values generated after log transformation, normalization (Percentile Shift, Scale or Normalize to control genes) and baseline transformation.
6. **Treatment of on-chip replicates:** Not Applicable.
7. **Flag values:** Not Applicable.
8. **Treatment of Control probes:** Not Applicable.
9. **Empty Cells:** Not Applicable.
10. **Sequence of events (CEL files):** The sequence of events involved in the processing of a CEL file is: Summarization→log transformation→Baseline Transformation.
11. **Sequence of events (CHP files):** If the data in the CHP file is already log transformed, then **GeneSpring GX** detects it and proceeds with the normalization step.

8.3 Guided Workflow steps

The *Guided Workflow* wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

Summary report (Step 1 of 8): The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized Expression values on the Y axis. An information message on the top of the wizard shows the number of samples and the sample processing details. By default, the *Guided Workflow* performs ExonRMA on the CORE probesets and Baseline Transformation to Median of all Samples. In case of CHP files, the defaults are Median Shift Normalization to 75 percentile and Baseline transformation to median of all samples. If the number of samples are more than 30, they are only represented in a tabular column. On clicking the *Next* button it will proceed to the next step and on clicking *Finish*, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. On doing a right click, the options of invert selection is displayed and on clicking the same the selection is inverted i.e., all the probes except the selected ones are highlighted in green. Figure 8.8 shows the Summary report with box-whisker plot.

Note: In the *Guided Workflow*, these default parameters cannot be changed. To choose different parameters use *Advanced Analysis*.

Experiment Grouping (Step 2 of 8): On clicking *Next*, the *Experiment Grouping* window appears which is the 2nd step in the **Guided Workflow**. It requires parameter values to be defined to group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the *Add parameter* button. Parameter values can be assigned by first selecting the

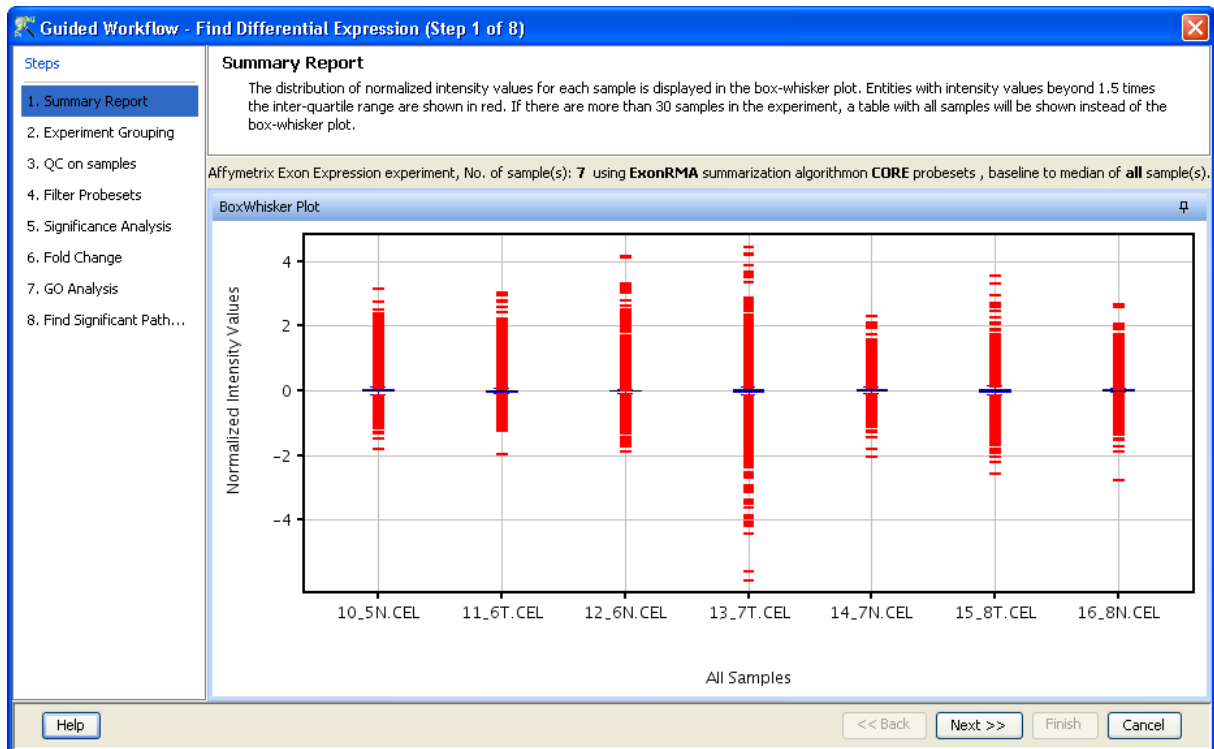




Figure 8.8: Summary Report

desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on **Clear**. Press **OK** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Guided Workflow**. The other parameters can be used in the **Advanced Analysis**.





Note: The *Guided Workflow* does not proceed further without grouping information.

Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file  icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

```
A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50
```

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file  icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click → *Properties* → *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click → *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

Note: The *Guided Workflow* by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the guided wizard.

Windows for experiment grouping and parameter editing are shown in figures 8.9 and 8.10 respectively.

Quality Control (Step 3 of 8): The 3rd step in the Guided Workflow is the QC on samples which displays three tiled windows when *CHP* files are used and four when *CEL* files are used as samples. They are as follows:

- Experiment grouping
- Hybridization Controls(only for CEL files)
- PCA scores
- Legend

See Figure 8.11 for more details.

The views in these windows are lassoed i.e., selecting the sample in any of the view highlights the sample in all the views.

The *Experiment Grouping* view shows the samples and the parameters present.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows

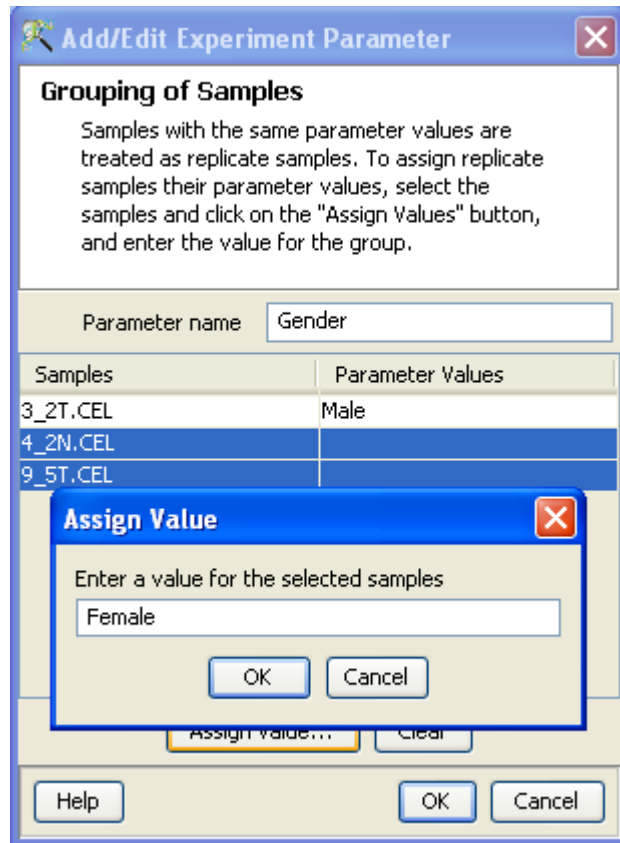


Figure 8.9: Experiment Grouping

viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed again on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

Filter probesets (Step 4 of 8): This operation removes by default, the lowest **20 percentile** of all the intensity values and generates a profile plot of filtered entities. This operation is performed on the raw signal values. The plot is generated using the normalized (not raw) signal values and samples grouped by the active interpretation. The plot can be customized via the right-click menu. This filtered Entity List will be saved in the Navigator window. The Navigator window can be viewed after exiting from *Guided Workflow*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Newer annotations can be added and existing ones removed using the *Configure Columns* button. Additional tabs in the

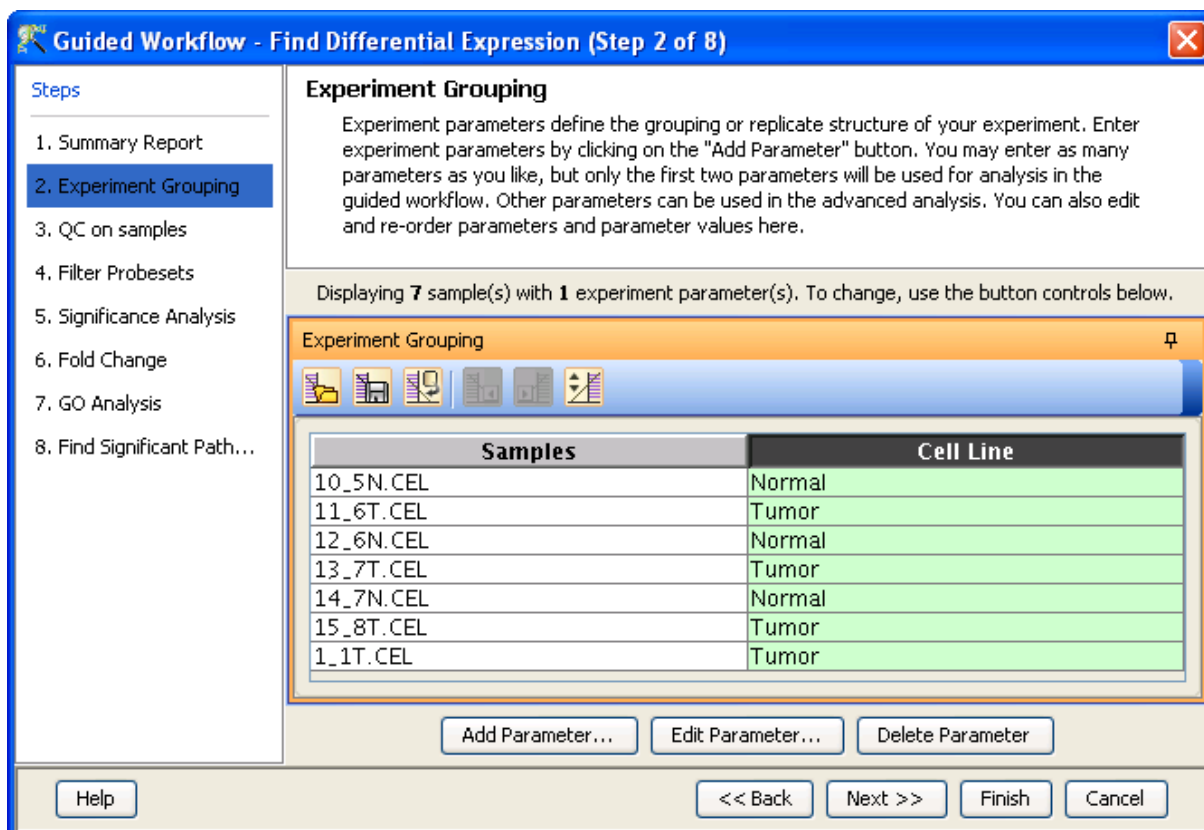


Figure 8.10: Edit or Delete of Parameters

Entity Inspector give the raw and the normalized values for that entity. The cutoff for filtering is set at 20 percentile and which can be changed using the button *Rerun Filter*. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. Figures 8.12 and 8.13 are displaying the profile plot obtained in situations having a single and two parameters. Re-run option window is shown in 8.14

Significance analysis(Step 5 of 8): Depending upon the experimental grouping, **GeneSpring GX** performs either T-test or ANOVA. The tables below describe broadly the type of statistical test performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I*, has 2 groups, the normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.
- **Example Sample Grouping II:** In this example, only one group, the tumor, is present. T-test against zero will be performed here.
- **Example Sample Grouping III:** When 3 groups are present (normal, tumor1 and tumor2) and one of the groups (tumor2 in this case) does not have replicates, statistical analysis cannot be performed. However if the condition tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.
- **Example Sample Grouping IV:** When there are 3 groups within an interpretation, One-way ANOVA will be performed.

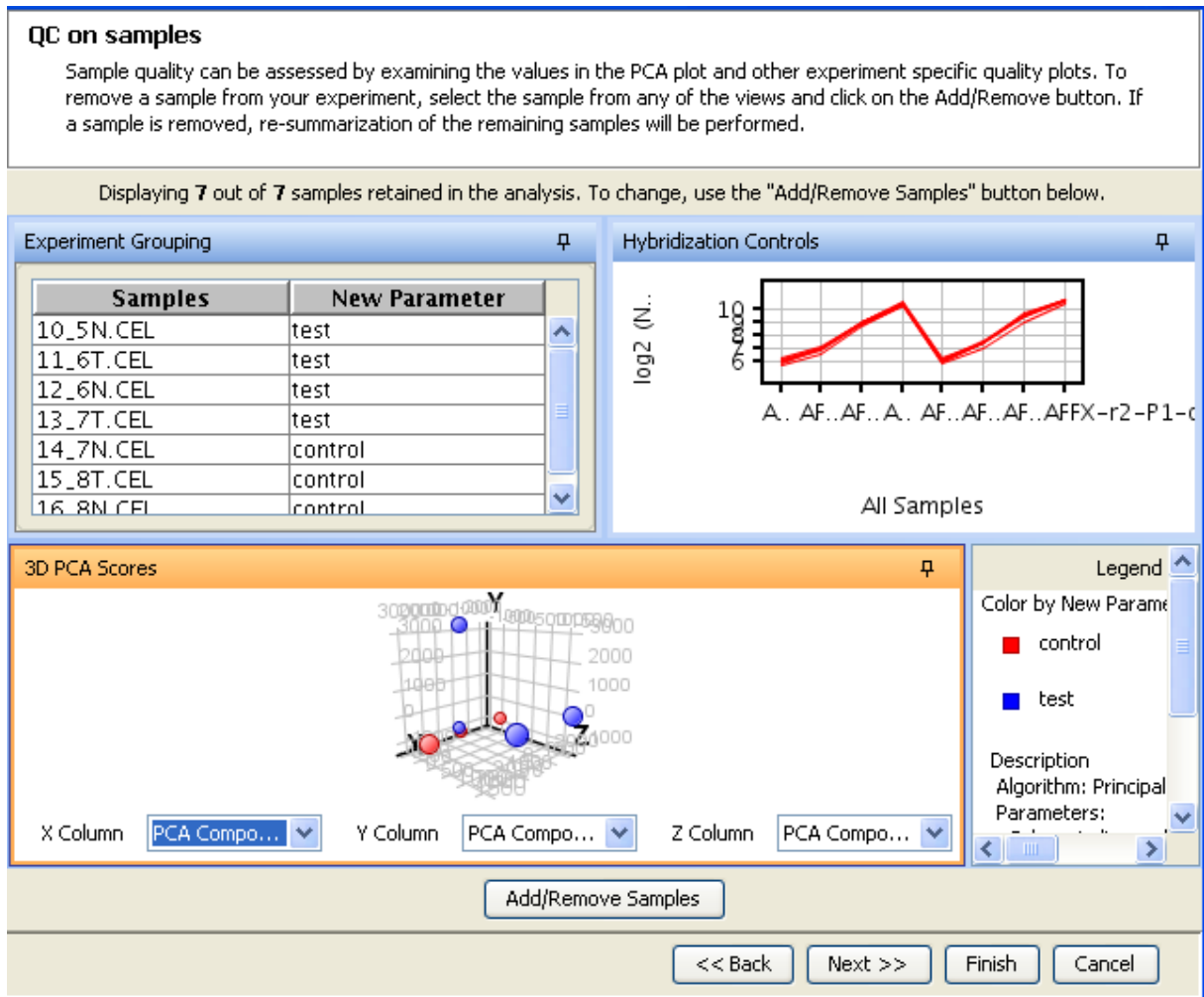


Figure 8.11: Quality Control on Samples

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 8.1: Sample Grouping and Significance Tests I

- **Example Sample Grouping V:** This table shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.
- **Example Sample Grouping VI:** In this table, a two-way ANOVA will be performed.
- **Example Sample Grouping VII:** In the example below, a two-way ANOVA will be performed

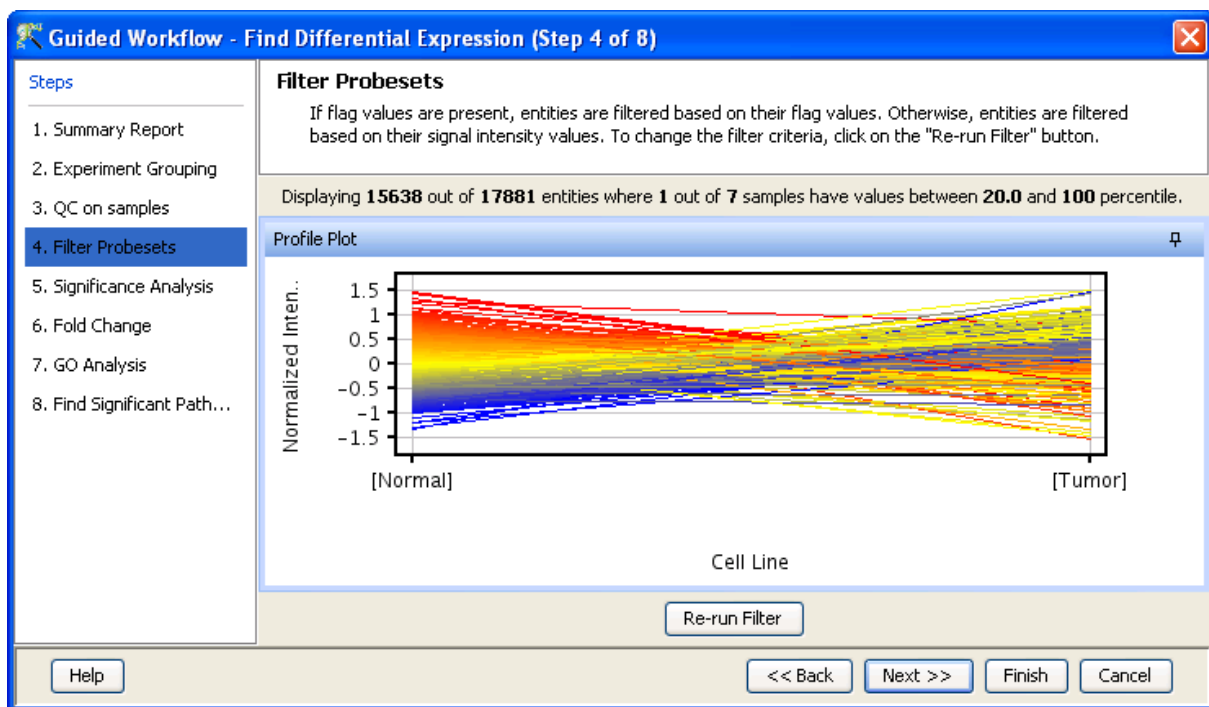


Figure 8.12: Filter Probesets-Single Parameter

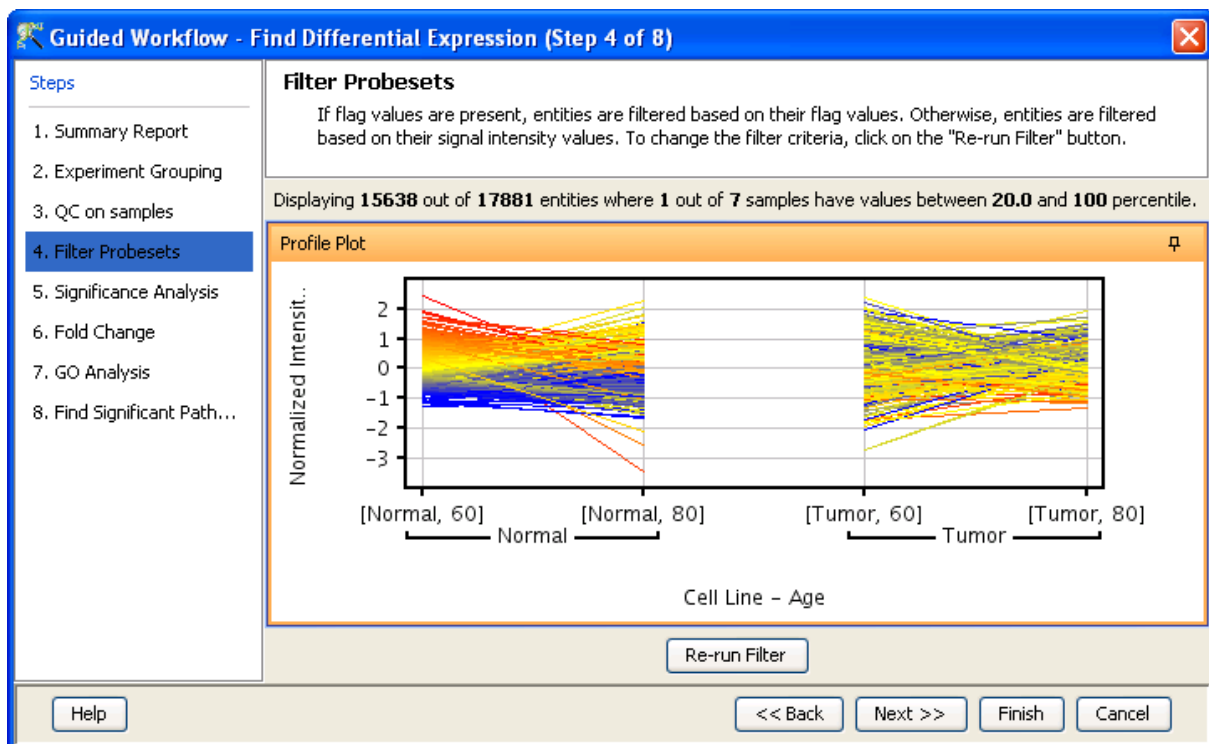


Figure 8.13: Filter Probesets-Two Parameters

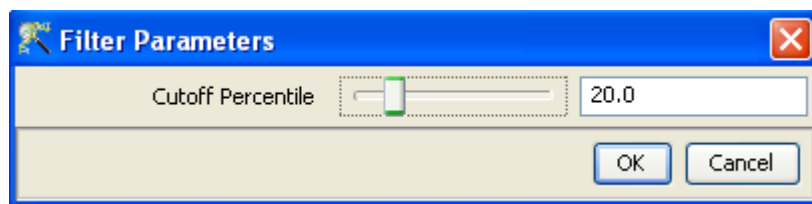


Figure 8.14: Rerun Filter

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 8.2: Sample Grouping and Significance Tests II

and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Statistical Tests: T-test and ANOVA

- **T-test: T-test unpaired** is chosen as a test of choice with a kind of experimental grouping shown in Table 1. Upon completion of T-test the results are displayed as three tiled windows.
 - A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *Regulation*.
 - *Differential expression analysis report* mentioning the Test description i.e. test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic* or *Permutative*).

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

- **Analysis of variance(ANOVA):** ANOVA is chosen as a test of choice under the experimental grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:
- A *p-value table* consisting of probe names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 8.3: Sample Grouping and Significance Tests III

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 8.4: Sample Grouping and Significance Tests IV

- *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and p-value computation type (*Asymptotic or Permutative*).
- *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

Special case: In situations when samples are not associated with at least one possible permutation of conditions (like Normal at 50 min and Tumor at 10 min mentioned above), no p-value can be computed and the **Guided Workflow** directly proceeds to **GO analysis**.

Fold-change (Step 6 of 8): Fold change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis. The wizard shows a table consisting of 3 columns: Probe Names, Fold change value and regulation (up or down). The regulation column depicts which one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using *Re-run Filter*. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1. A profile plot is also generated. Upregulated entities are shown in red. The color can be changed using the Right-click → *Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cutoff in the experiment Navigator.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 8.5: Sample Grouping and Significance Tests V

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 8.6: Sample Grouping and Significance Tests VI

Note: Fold Change step is skipped and the *Guided Workflow* proceeds to the *GO Analysis* in case of experiments having 2 parameters.

Fold Change view with the spreadsheet and the profile plot is shown in Figure 8.17.

Gene Ontology analysis(Step 7 of 8): The *GO Consortium* maintains a database of controlled vocabularies for the description of molecular function, biological process and cellular location of gene products. The GO terms are displayed in the Gene Ontology column with associated *Gene Ontology Accession* numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

The GO analysis wizard shows two tabs comprising of a spreadsheet and a *GO tree*. The *GO Spreadsheet* shows the *GO Accession* and *GO terms* of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However, the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.1 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 8.7: Sample Grouping and Significance Tests VII

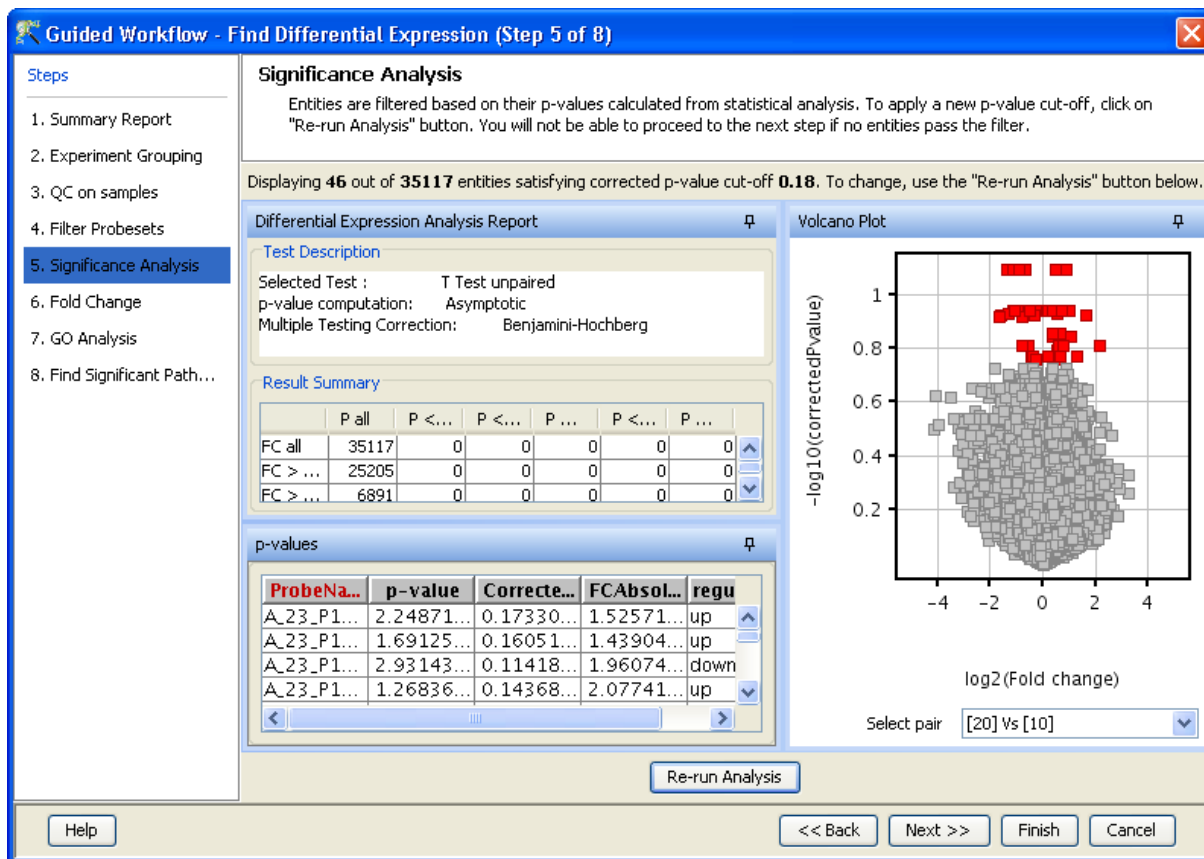


Figure 8.15: Significance Analysis-T Test

of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button

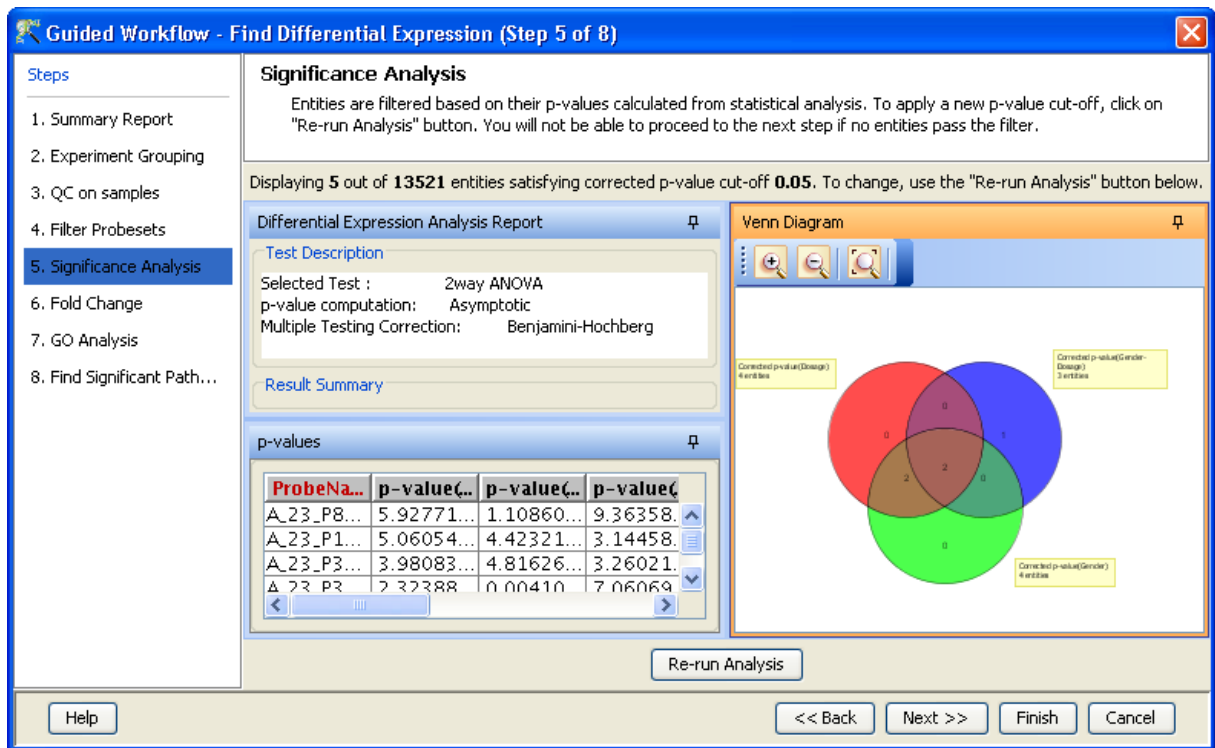


Figure 8.16: Significance Analysis-Anova

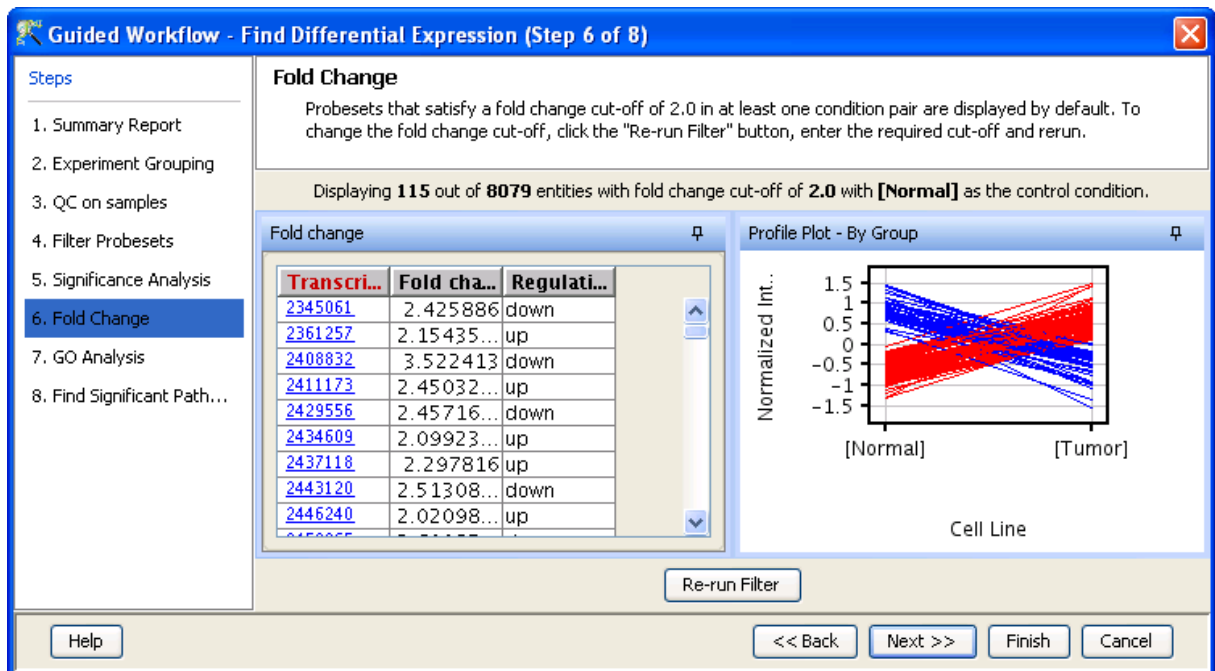


Figure 8.17: Fold Change

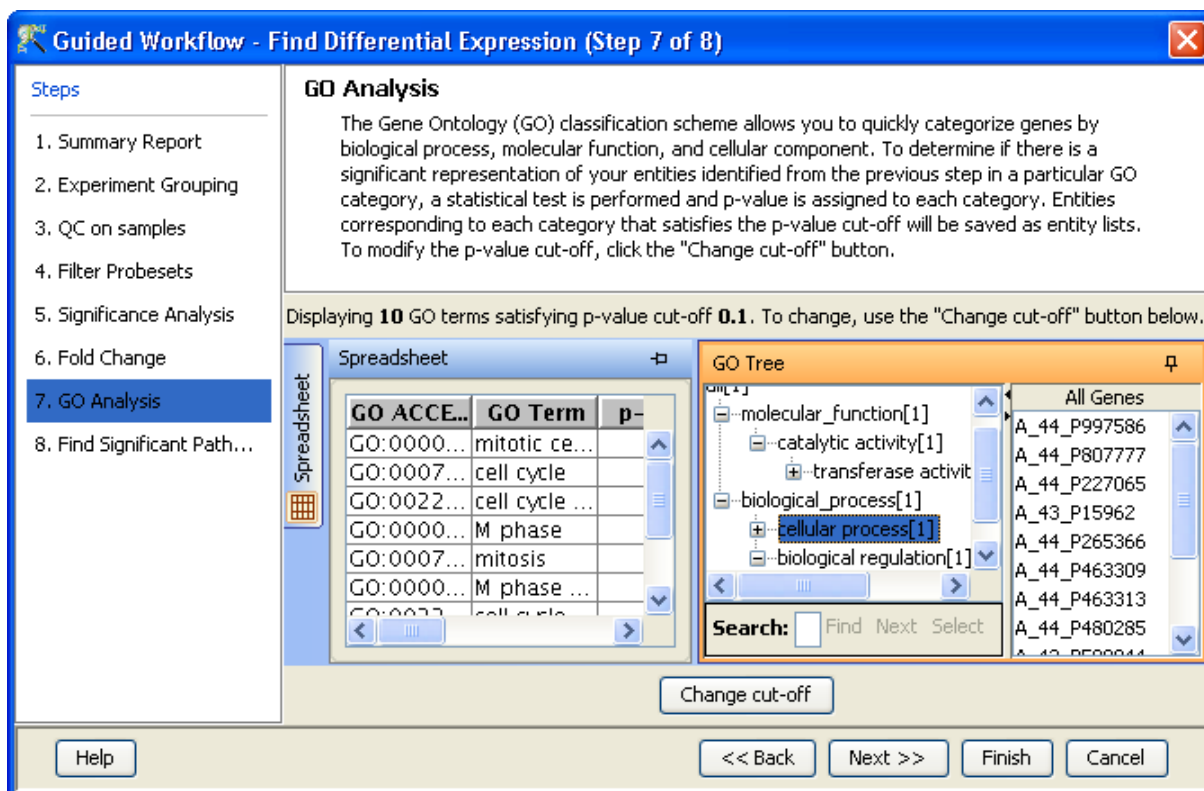


Figure 8.18: GO Analysis

is also provided on the GO tree panel to search using some keywords

Note : In **GeneSpring GX** GO analysis implementation, all the three component: Molecular Function, Biological Processes and Cellular location are considered together.

On finishing the GO analysis, the *Advanced Workflow* view appears and further analysis can be carried out by the user. At any step in the Guided workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any) and the *Advanced Workflow* view appears.

Find Significant Pathways (Step 8 of 8): This step in the Guided Workflow finds relevant pathways from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the Fold Change (step 6 of 8). This view shows two tables-

- The Significant Pathways table shows the names of the pathways as well as the number of nodes and entities in the pathway and the p-values. It also shows the number of entities that are similar to the pathway and the entity list. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.
- The Non-significant Pathways table shows the pathways in the tool that do not have a single entity in common with the ones in the given entity list.

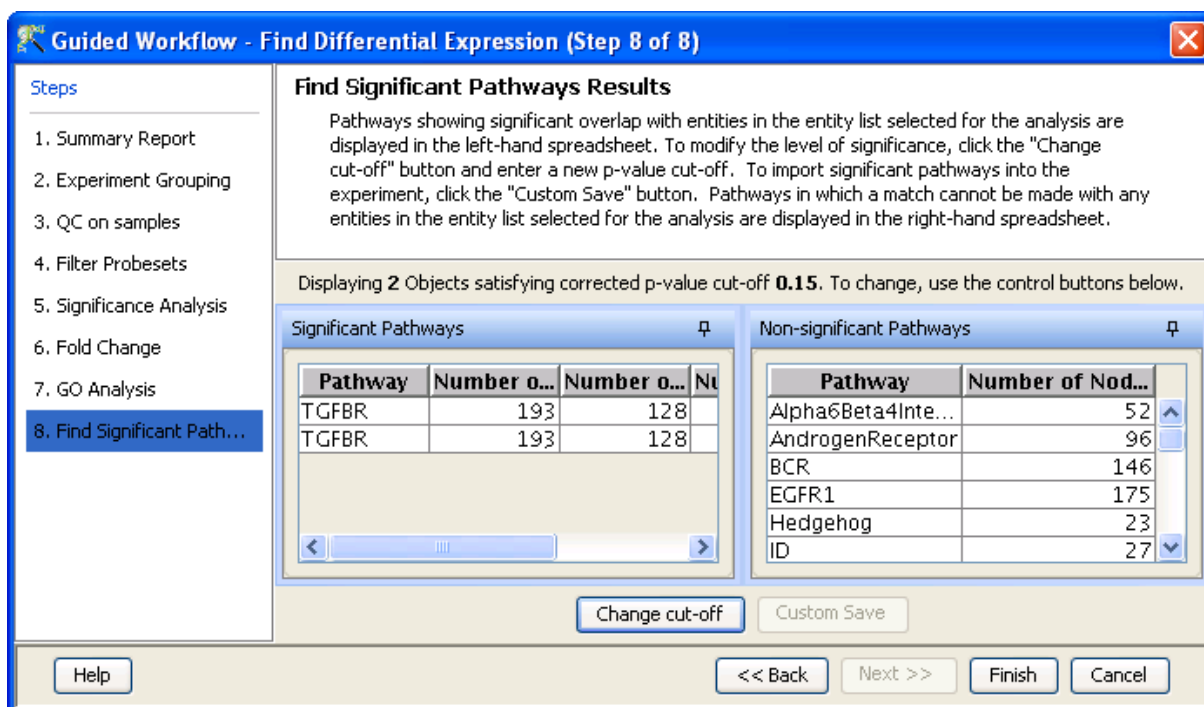


Figure 8.19: Find Significant Pathways

The user has an option of changing the p-value cut-off (using *Change cutoff*) and also to save specific pathways using the *Custom Save* option. On clicking, *Finish* the main tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of the Guided Workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in the Guided Workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any). See figure 8.19.

Note: In case the user is using **GeneSpring GX** for the first time, this option will give results using the demo pathways. The user can upload the pathways of his/her choice by using the option *Import BioPax pathways* under **Tools** in the **Menu** bar. Later instead of reverting to the Guided Workflow the user can use the option *Find Significant Pathways* in **Results Interpretation** under the same Workflow.

The default parameters used in the Guided Workflow is summarized below

8.4 Advanced Workflow

The *Advanced Workflow* offers a variety of choices to the user for the analysis. Several different summarization algorithms are available for probeset summarization. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment

	Parameters	Parameter values
Expression Data Transformation	Thresholding	1.0
	Normalization	Quantile
	Baseline Transformation	Median to all samples
	Summarization	RMA16
Filter by		
1.Flags	Flags Retained	Not Applicable
2.Expression Values	(i) Upper Percentile cutoff	100
	(ii) Lower Percentile cutoff	20
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1

Table 8.8: Table of Default parameters for Guided Workflow

using the *Advanced Workflow*, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Advanced. Clicking **OK** will open a New Experiment Wizard, which then proceeds as follows:

8.4.1 Creating an Affymetrix ExonExpression Experiment

An *Advanced Workflow* Analysis can be done using either CEL or CHP files. However, a combination of both file types cannot be used. Only transcript summarized CHP files can be loaded in a project.

New Experiment (Step 1 of 7): Load data As in case of *Guided Workflow*, either data files can be imported or else pre-created samples can be used.

- For loading new CEL/CHP files, use *Choose Files*.
- If the CEL/CHP files have been previously used in experiments *Choose Samples* can be used.

Step 1 of 7 of Experiment Creation, the **Load Data** window, is shown in Figure 8.20.

New Experiment (Step 2 of 7): Selecting ARR files ARR files are Affymetrix files that hold annotation information for each sample CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample. Click on **Next** to proceed to the next step.

Step 2 of 7 of Experiment Creation, the Select ARR files window, is depicted in the Figure 8.21.

New Experiment (Step 3 of 7): Pairing of transcript and probeset level files This step of the wizard is used in the case of Affymetrix Exon Splicing experiment type.

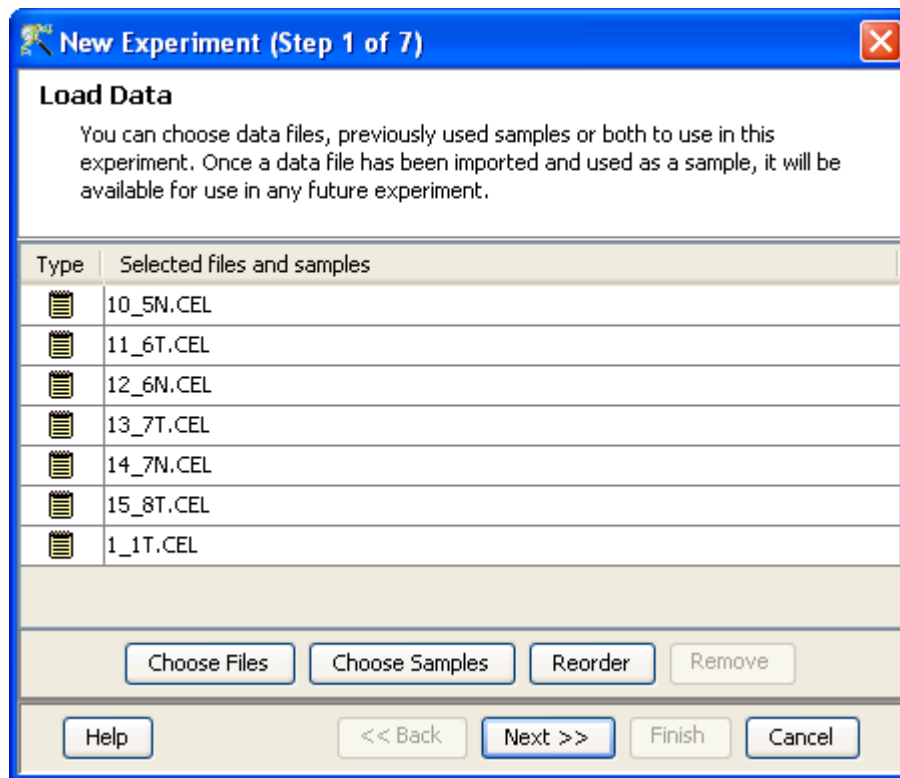


Figure 8.20: Load Data

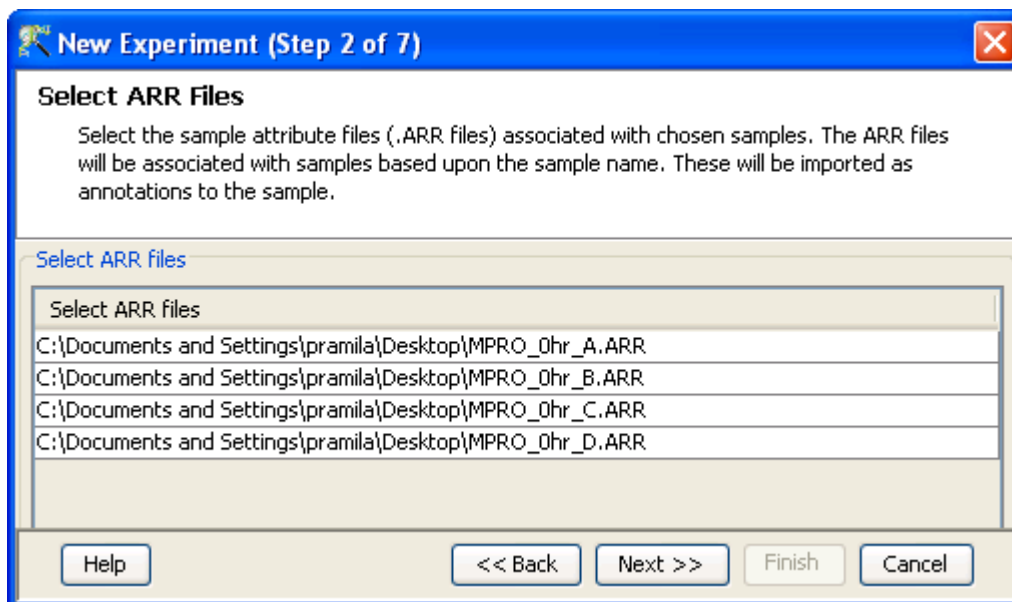


Figure 8.21: Select ARR files

New Experiment (Step 4 of 7): Preprocess Baseline Options This step is specific for CEL files. Any one of the Summarization algorithms provided from the drop down menu can be chosen to summarize the data. The available summarization algorithms are:

- The RMA Irazarry *et al.* [Ir1, Ir2, Bo].
- The PLIER16 Hubbell [Hu2].
- The IterativePLIER16

Subsequent to probeset summarization, baseline transformation of the data can be performed. The baseline options include:

- ***Do not perform baseline***
- ***Baseline to median of all samples:*** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- ***Baseline to median of control samples:*** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

This step also enables the user to select the meta-probeset list, using which the summarization is done.

Three metaprobeset lists (sourced from Expression Console by Affymetrix) are pre-packaged with the data library file for the corresponding ExonChip. They are called the Core, Extended and Full.

1. The Core list comprises 17,800 transcript clusters from RefSeq and full-length GenBank mRNAs.
2. The Extended list comprises 129k transcript clusters including cDNA transcripts, syntenic rat and mouse mRNA, and Ensembl, microRNA, Mitomap, Vegagene and VegaPseudogene annotations.
3. The full list comprises 262k transcript clusters including ab-initio predictions from Geneid, Genscan, GENSCAN Suboptimal, Exoniphy, RNAGene, SgpGene and TWINSCAN.

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar. Figure 8.22 shows the Step 4 of 7 of Experiment Creation.

New Experiment (Step 5 of 7): This step is specific for CHP files only. See Figure 8.23 It gives the user the following normalization options

- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.

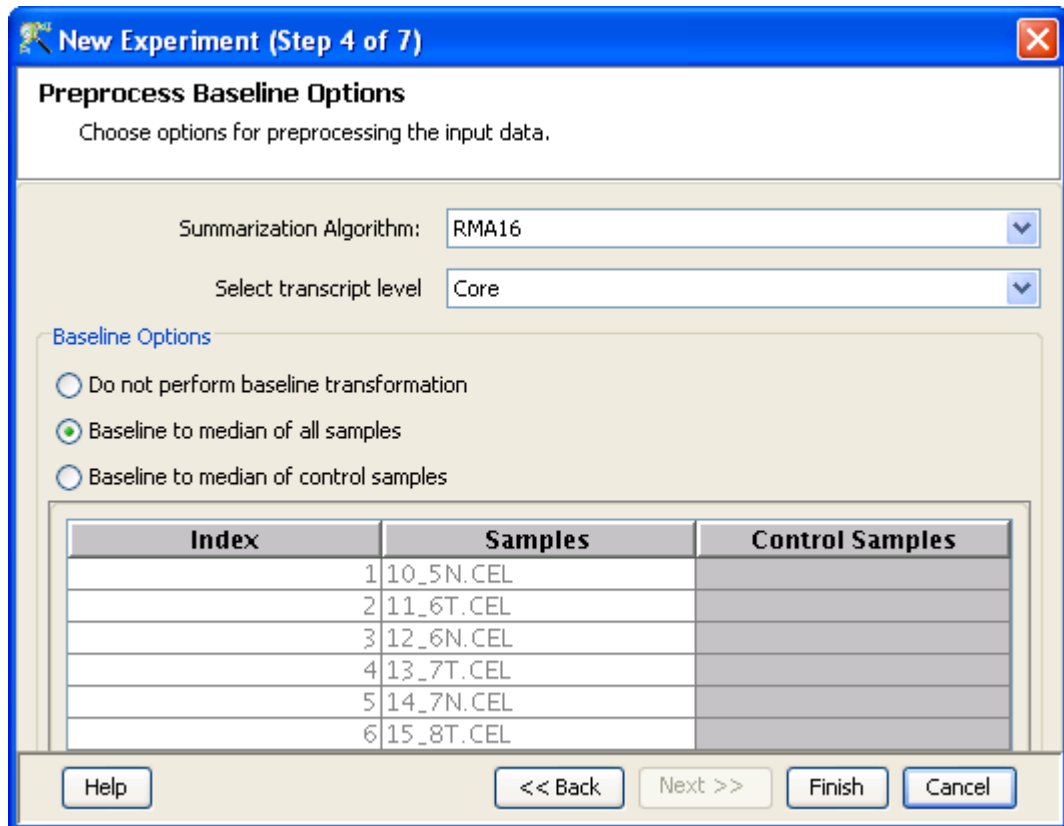


Figure 8.22: Summarization Algorithm

- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the 'Assign Value' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.

For details on the above normalization methods, refer to section on [Normalization Algorithms](#).

New Experiment (Step 6 of 7): If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the

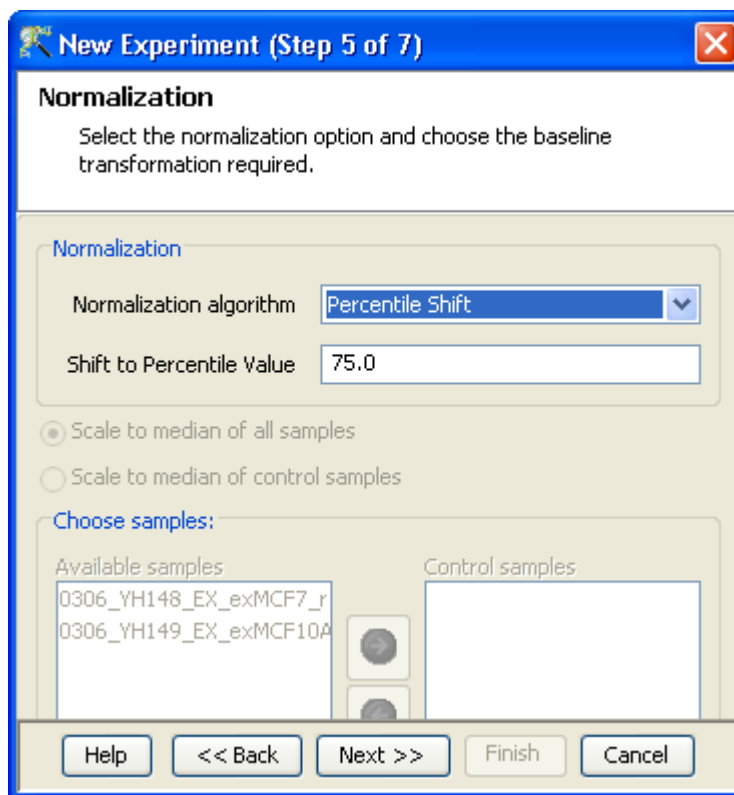


Figure 8.23: Normalization

technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard. See figures 8.24, 8.25 and 8.26.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

New Experiment (Step 7 of 7): This step allows the user to perform baseline transformation. The methods available are the same as those used for CEL files in Step 4 of 7.

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to *View* in Toolbar. The final step of Experiment Creation (CHP file specific) is shown in Figure 8.27.

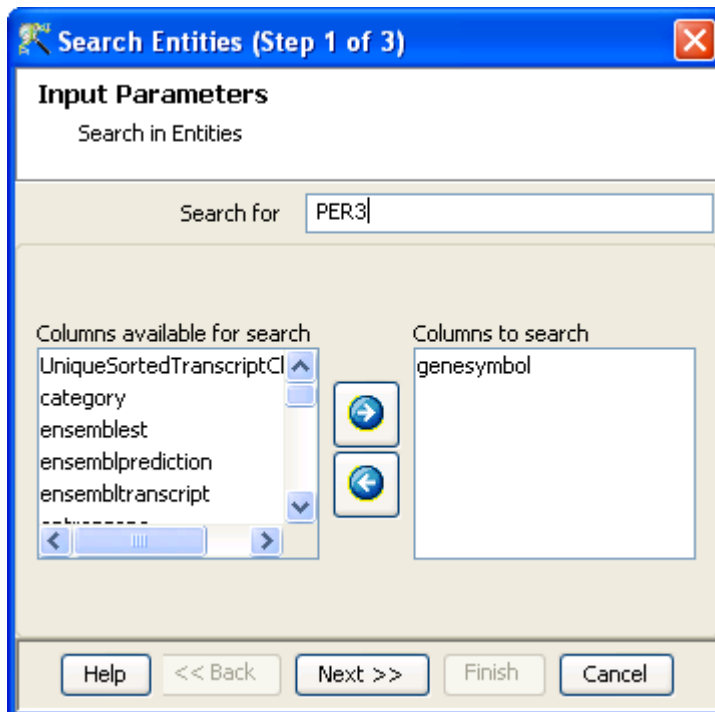


Figure 8.24: Search entities

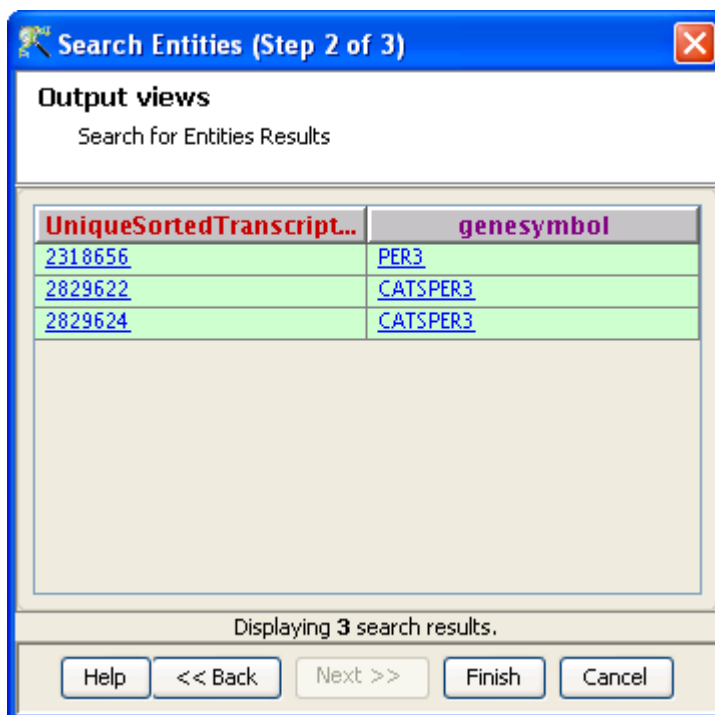


Figure 8.25: Output Views

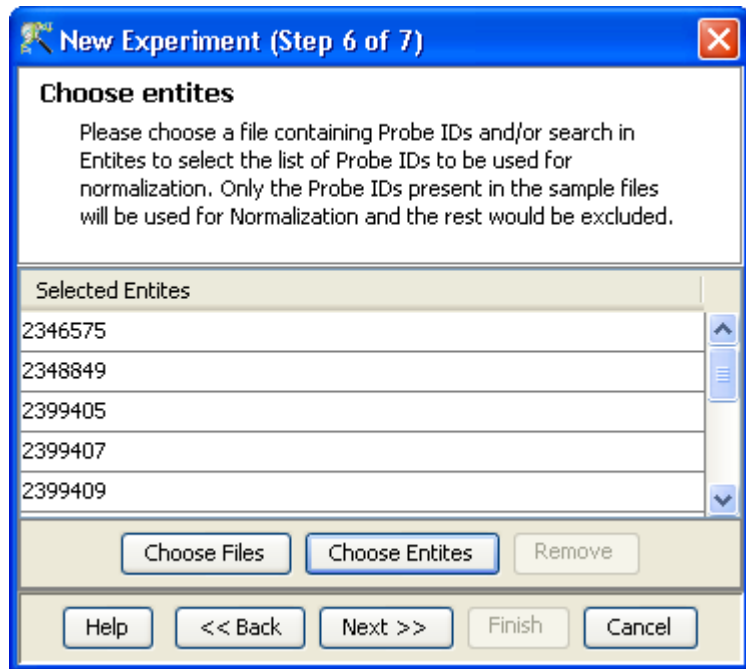


Figure 8.26: Choose Entities

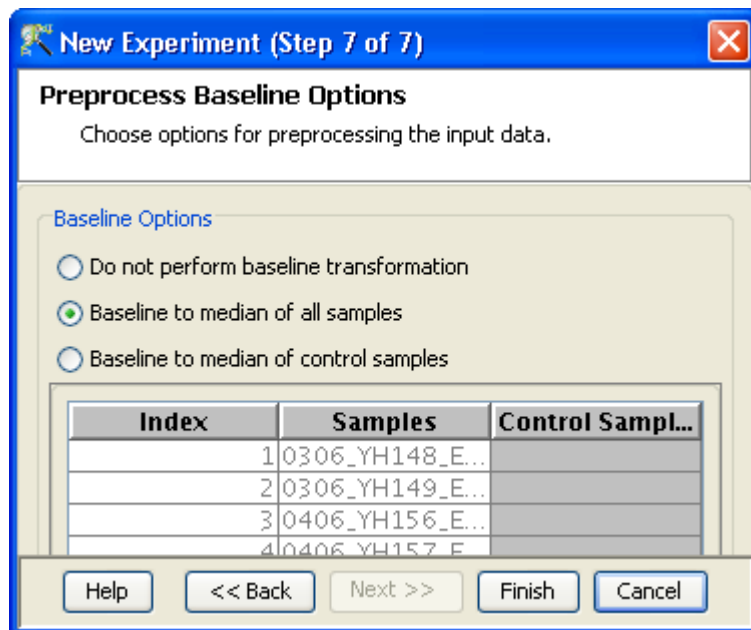


Figure 8.27: Normalization and Baseline Transformation

8.4.2 Experiment setup

- **Quick Start Guide**

Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the *Advanced Workflow*, the method of analysis, the details of the algorithms used and the interpretation of results.

- **Experiment Grouping:** *Experiment parameters* defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

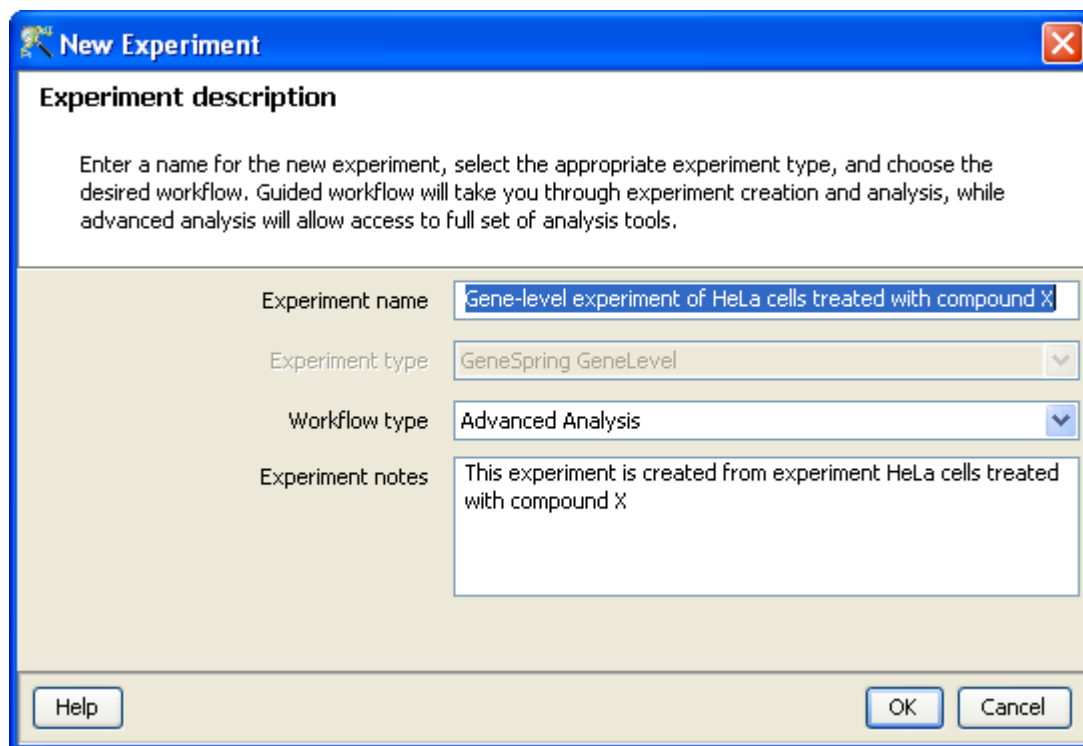


Figure 8.28: Gene Level Experiment Creation

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

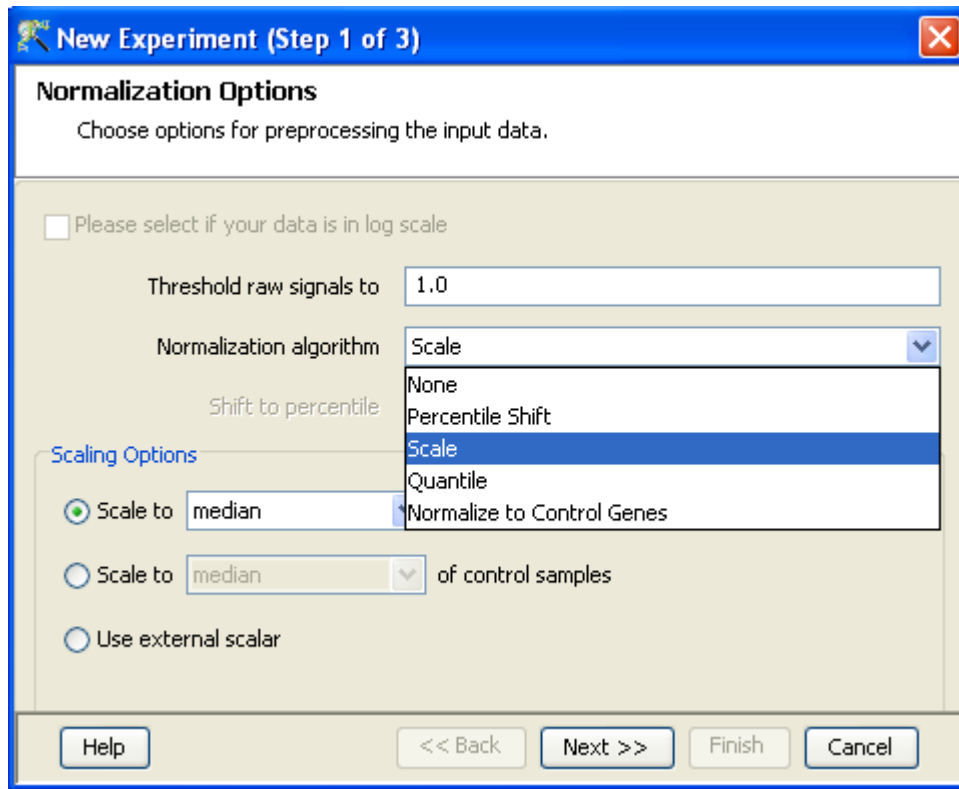


Figure 8.29: Gene Level Experiment Creation - Normalization Options

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

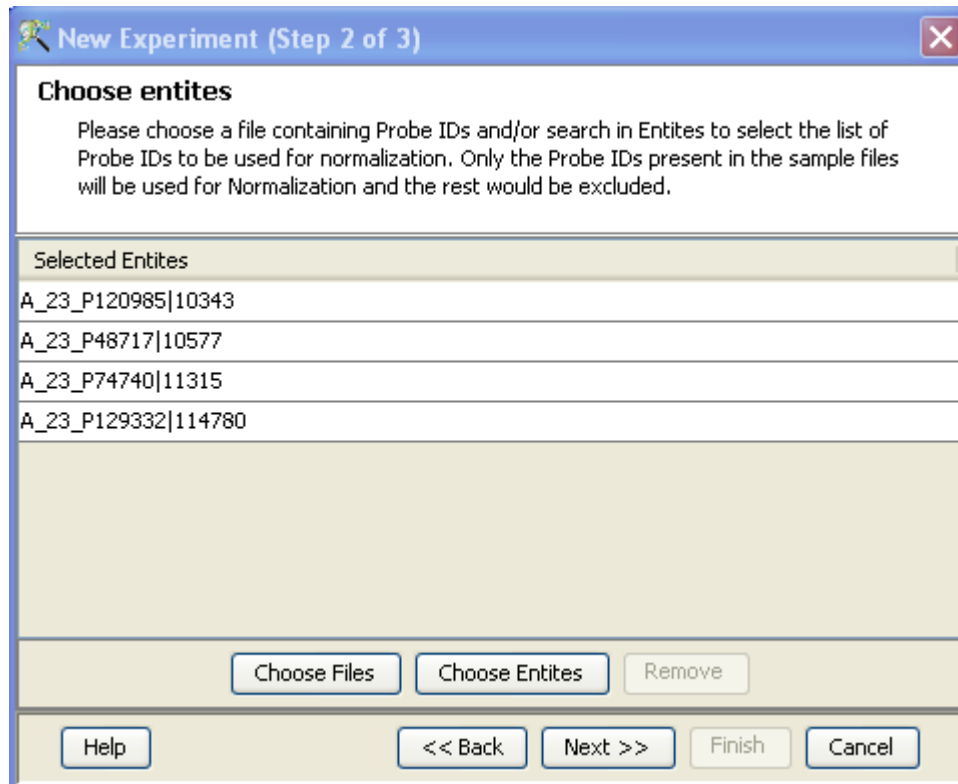


Figure 8.30: Gene Level Experiment Creation - Choose Entities

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

8.4.3 Quality Control

- **Quality Control on Samples**

Quality Control or QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows three tiled windows:

- Experiment grouping and hybridization controls (applicable for CEL files).
- 3D PCA scores, Correlation coefficients and Correlation plot tabs.
- Legend.

Figure 9.13 has the 3 tiled windows which reflect the QC on samples.

Experiment Grouping shows the parameters and parameter values for each sample.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the

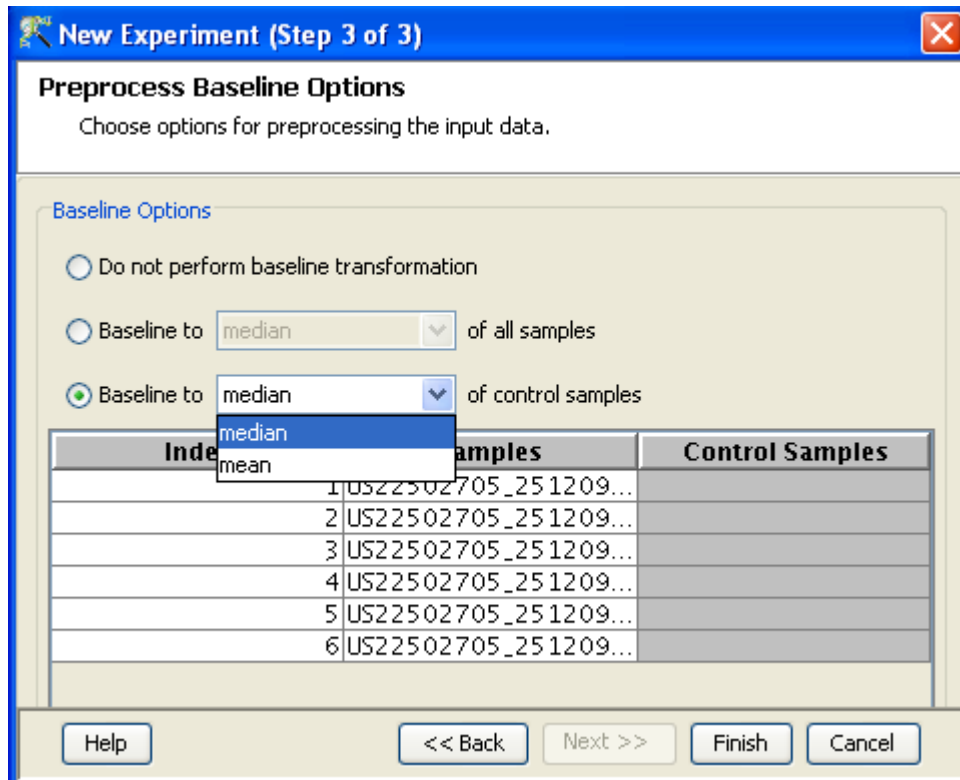


Figure 8.31: Gene Level Experiment Creation - Preprocess Baseline Options

hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$

The heatmap is colorable by Experiment Factor information via Right-Click → Properties. The intensity levels in the heatmap can also be customized here.

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster

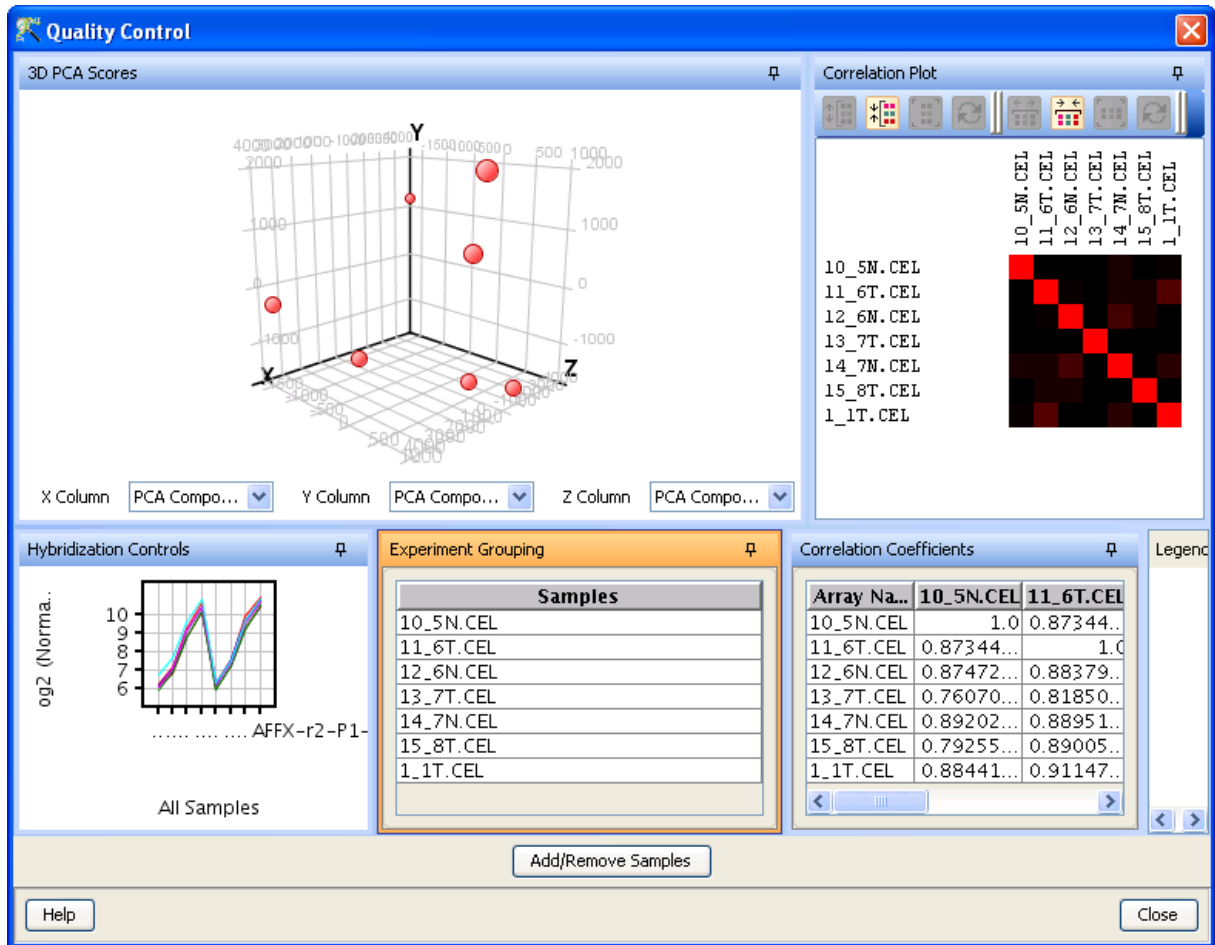


Figure 8.32: Quality Control

together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click** → **Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-summarization of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags** No flags are generated during creation of exon expression experiment.
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

8.4.4 Analysis

- **Statistical Analysis**

For details refer to section [Statistical Analysis](#) in the advanced workflow.

- **Filter on Volcano Plot**

For details refer to section [Filter on Volcano Plot](#)

- **Fold Change**

For details refer to section [Fold Change](#)

- **Clustering**

For details refer to section [Clustering](#)

- **Find Similar Entities**

For details refer to section [Find Similar Entities](#)

- **Filter on Parameters**

For details refer to section [Filter on Parameters](#)

- **Principal Component Analysis**

For details refer to section [PCA](#)

8.4.5 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)

- **Run Prediction** For details refer to section [Run Prediction](#)

8.4.6 Results

- **Gene Ontology (GO) analysis**

GO is discussed in a separate chapter called [Gene Ontology Analysis](#).

- **Gene Set Enrichment Analysis (GSEA)**

Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).

- **Gene Set Analysis (GSA)**

Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).

- **Pathway Analysis**

Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).

- **Find Similar Entity Lists**

This feature is discussed in a separate section called [Find Similar Entity Lists](#)

- **Find Significant Pathways**

This feature is discussed in a separate section called [Find Significant Pathways](#).

- **Launch IPA**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Import IPA Entity List**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Extract Interactions via NLP**

This feature is discussed in detail in the chapter [Pathway Analysis](#).

8.4.7 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)

- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)

- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)

- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

8.4.8 Algorithm Technical Details

Here are some technical details of the Exon RMA16, Exon PLIER16, and Exon IterPLIER16 algorithms.

Exon RMA 16. Exon RMA does a GC based background correction (described below and performed only with the PM-GCBG option) followed by Quantile normalization followed by a Median Polish probe summarization, followed by a Variance Stabilization of 16. The computation takes roughly 30 seconds per CEL file with the *Full* option.

GCBG background correction bins background probes into 25 categories based on their GC value and corrects each PM by the median background value in its GC bin. RMA does not have any configurable parameters.

Exon PLIER 16. Exon PLIER does Quantile normalization followed by the PLIER summarization using the PM or the PM-GCBG options, followed by a Variance Stabilization of 16. The PLIER implementation and default parameters are those used in the Affymetrix Exact 1.2 package. PLIER param-

ters can be configured from *Tools* \rightarrow *Options* \rightarrow *Affymetrix Exon Summarization Algorithms* \rightarrow *Exon PLIER/IterPLIER*.

Exon IterPLIER 16. Exon IterPLIER does Quantile normalization followed by the IterPLIER summarization using the PM or the PM-GCBG options, followed by a Variance Stabilization of 16. IterPLIER runs PLIER multiple times, each time with a smaller subset of the probes obtained by removing outliers from the previous PLIER run. IterPLIER parameters can be configured from *Tools* \rightarrow *Options* \rightarrow *Affymetrix Exon Summarization Algorithms* \rightarrow *Exon PLIER/IterPLIER*.

Chapter 9

Analyzing Affymetrix Exon Splicing Data

Alternative splicing is defined as variations in RNA splicing mechanisms resulting in multiple splice variants, each specific to a stage or condition of the cell. Affymetrix Exon chips are used for studying the alternative splicing of genes. A large population of human mRNAs undergo alternative splicing which generates splice variants that produce proteins with distinct and sometimes even antagonistic functions. Also changes in splicing signals or in sequences regulating splicing have been implicated as the cause for certain genetic mutations which result in human diseases. Thus measuring changes in splicing patterns is integral to understanding the disease mechanism or biological process under study. **GeneSpring GX** supports Exon Splicing analysis using the Affymetrix Exon Arrays.

9.1 Running the Affymetrix Exon Splicing Workflow

Upon launching **GeneSpring GX** , the startup is displayed with 3 options.

1. Create new project
2. Open existing project
3. Open recent project

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting *Create new project*, a window appears in which details (Name of the project and Notes) can be recorded. Press **OK** to proceed. An Experiment Selection Dialog window then appears with two options

1. Create new experiment
2. Open existing experiment

Selecting *Create new experiment* allows the user to create a new experiment (steps described below). *Open existing experiment* allows the user to use existing experiments from any previous projects in the current project. Choosing *Create new experiment* opens up a **New Experiment dialog** in which experiment name can be assigned. The experiment type should then be specified. The drop-down menu gives the user the option to choose between the Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Real Time PCR, Pathway, Generic Single Color and Two Color experiment types. The **Advanced Workflow** is the only option for the Affymetrix Exon Splicing experiment.

Upon clicking *OK*, the Affymetrix Exon Splicing experiment creation wizard appears.

9.1.1 Creating an Affymetrix Exon Splicing Experiment

An **Advanced Workflow** analysis can be done using either CEL or CHP files. However, a combination of both file types cannot be used. If CHP files are being used for analysis, then both transcript (gene) summarized and probeset (exon) summarized files need to be present for a sample.

New Experiment (Step 1 of 7): Load data An experiment can be created either using data files or using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. A sample, on the other hand is created within the tool, when it associates the data file with its appropriate technology. For more details, refer to the section on [Technology](#). Thus a sample created within a technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology.

- For loading new CEL/CHP files, use *Choose Files*.
- If the CEL/CHP files have been previously used in experiments *Choose Samples* can be used.

Note: In **GeneSpring GX Exon Splicing Workflow**, experiment creation using CHP files requires 2 types of CHP files per array i.e., the transcript level CHP file and the probeset level CHP file. This is necessary as **GeneSpring GX** requires the probeset level data for splicing analysis. If the user has not provided the required files for each array, **GeneSpring GX** prompts the user to provide the necessary files. Refer Figure 9.2. Additionally if the same experiment is created again from *Project Navigator* → *Experiment Name* → *Right click* → *Create New Experiment*, the files taken into account are only the transcript level files and the user needs to provide the probeset level files. These files can be loaded from the *Choose Samples* option.

Step 1 of 7 of Experiment Creation, the **Load Data** window, is shown in figure 9.1.

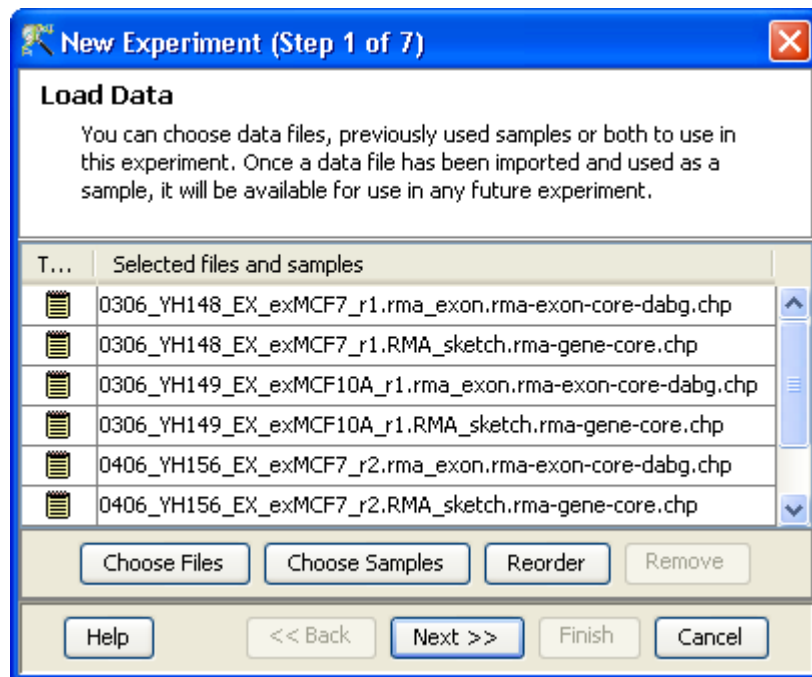


Figure 9.1: Load Data

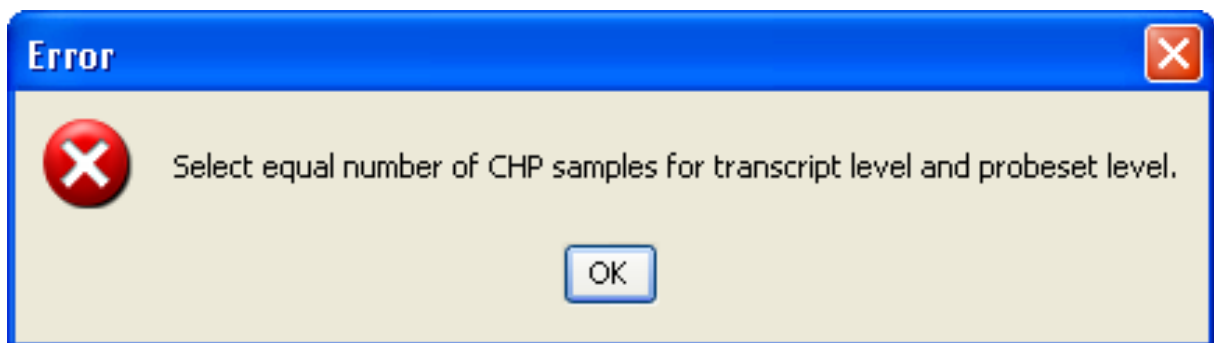


Figure 9.2: Error Message

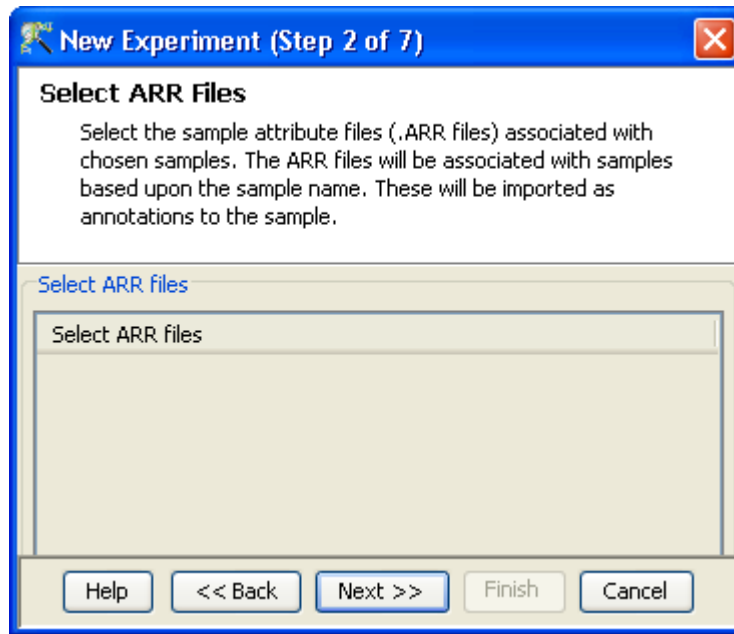


Figure 9.3: Select ARR files

New Experiment (Step 2 of 7): Selecting ARR files ARR files are Affymetrix files that hold annotation information for each sample's CEL and CHP file and are associated with the sample based on the sample name. These are imported as annotations to the sample. Click on *Next* to proceed to the next step.

Step 2 of 7 of Experiment Creation, the **Select ARR files** window, is depicted in the figure 9.3.

New Experiment (Step 3 of 7): Pairing of transcript and probeset level files This step is specific to CHP files. The tool pairs both the CHP files of a sample automatically, based on the file names. But in case the naming is different and the pairing done is incorrect, the user can change the pairing by selecting the file and moving it with the help of the buttons provided on the right side of the wizard. See figure 9.4.

New Experiment (Step 4 of 7): Preprocess Baseline Options Specific to the CEL files, step 4 provides three summarization algorithms. The suffix 16 in these algorithms denotes a variance stabilization addition of 16 to the result of each algorithm.

- RMA16 Irazarry *et al.* [Ir1, Ir2, Bo].
- PLIER16 Hubbell [Hu2].
- IterativePLIER16

The meta-probe set and the probe set list using which the summarization is done on transcript and probeset level respectively, is also chosen at this step.

The three meta-probe set and probe set lists, namely core, extended and full (sourced from Expression Console by Affymetrix) are pre-packaged with the data library file for the corresponding ExonChip.

Details of the meta probeset lists are given below. For more details on the same, refer to http://www.affymetrix.com/support/technical/whitepapers/exon_genesummary_whitepaper.pdf

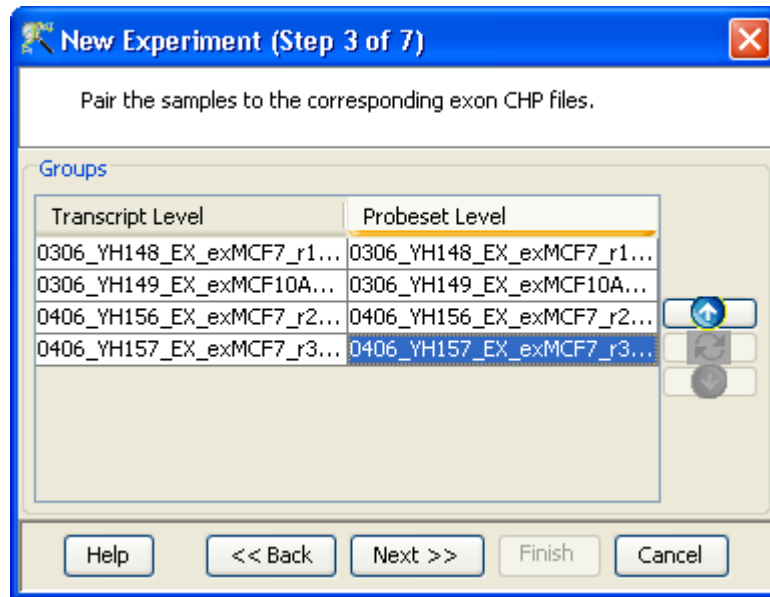


Figure 9.4: Pairing of CHP files

1. The Core meta-probe set list comprising 17,800 transcript clusters groups unique exon level probe sets with strong annotation support e.g., from RefSeq and other GenBank alignments of 'complete CDS' transcripts, into transcript clusters.
2. The Extended list comprising 129K transcript clusters groups unique exon level probe sets with empirical annotation support into transcript clusters. This includes cDNA transcripts, mapped syntenic mRNA from rat and mouse, and probe sets which are part of the Ensembl or Vega transcript annotation set.
3. The full list which groups all unique exon level probe sets comprises 262K transcript clusters including ab-initio predictions from Geneid, Genscan, GENSCAN Suboptimal, Exoniphy, RNA-gene, SgpGene and TWINSCAN.

Probe sets are graded according to the highest confidence evidence supporting it. Details of the probeset lists are given below:

1. The probes targeting exons with RefSeq mRNA evidence are regarded as the most confident and are present in the core probeset list. Core probe sets are supported with the most reliable evidence.
2. The probes targeting exons with EST evidence are referred to as "Extended" probes and are present in the extended probeset list.
3. The probes targeting putative computational exon predictions have the least confidence and are present in the full list.

The full list includes both the core and extended lists while the extended list contains the core probeset list.

Subsequent to probeset summarization, Baseline Transformation of the data can be performed. Baseline Transformation is carried out row-wise across all samples. This data processing step is particu-

larly useful when visualizing the results in a profile plot or heat map. The baseline transformation options, available in **GeneSpring GX** are:

- ***Do not perform baseline***
- ***Baseline to median of all samples:*** For each row (probe), the median of the log summarized values across all the samples is calculated. This value is then subtracted from the probe value for all samples.
- ***Baseline to median of control samples:*** Here control samples are used to calculate the median value for each probe. This value is then subtracted from the probe value for all samples. The controls could be an individual control for each sample or it could be a set of controls. Alternatively, a set of samples can be used as controls for all samples. For specifying the control for a sample, select the sample and click on ***Assign value***. This opens up the ***Choose Control Samples*** window from where the samples designated as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. Click on ***Ok***. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Clicking ***Finish*** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to ***View*** in Toolbar. Figure 9.5 shows the Step 4 of 7 of Experiment Creation.

New Experiment (Step 5 of 7): This step is specific for CHP files only. It gives the user the following normalization options. See figure 9.6.

- ***Percentile Shift:*** On selecting this normalization method, the ***Shift to Percentile Value*** box gets enabled allowing the user to enter a specific percentile value using which normalization is performed.
- ***Scale:*** On selecting this normalization method, an option is presented to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the ***Available Samples*** in the ***Choose Samples*** box. The ***Shift to percentile*** box is disabled and the percentile is set at a default value of 50.
- ***Normalize to control genes:*** After selecting this option, the user has to specify the control genes in the next wizard. The median of the control genes is then used for normalization.
- ***Normalize to External Value:*** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the '*Assign Value*' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.

For details on the above normalization methods, refer to section [normalization](#)

New Experiment (Step 6 of 7): If the ***Normalize to control genes*** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.

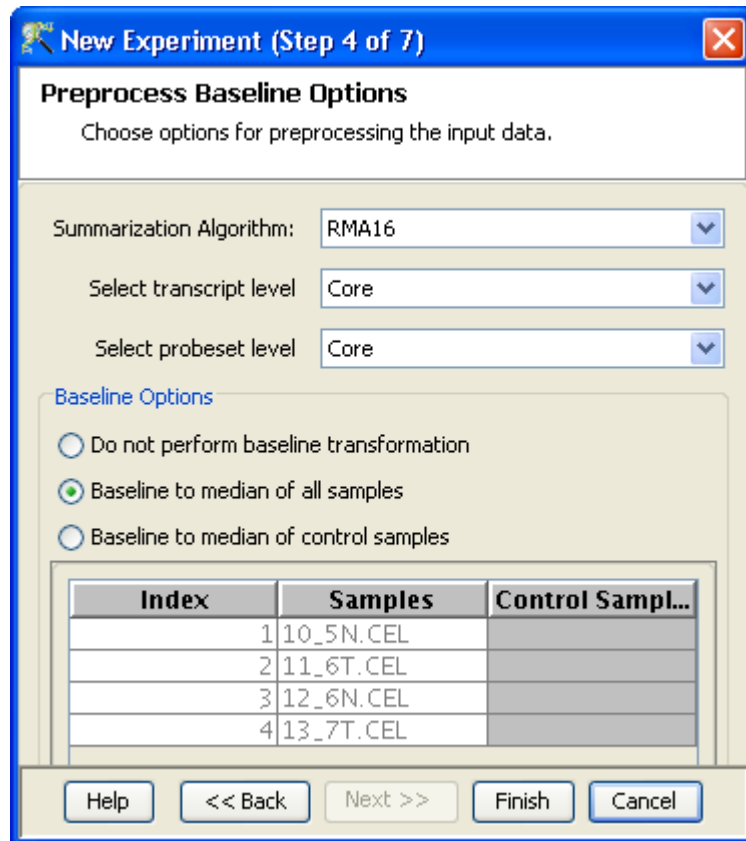


Figure 9.5: Summarization Algorithm

- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector. See figure 9.7.

New Experiment (Step 7 of 7): This step allows the user to perform baseline transformation. The methods available are the same as those used for CEL files in Step 4 of 7.

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to *View* in Toolbar. The final step of Experiment Creation (CHP file specific) is shown in

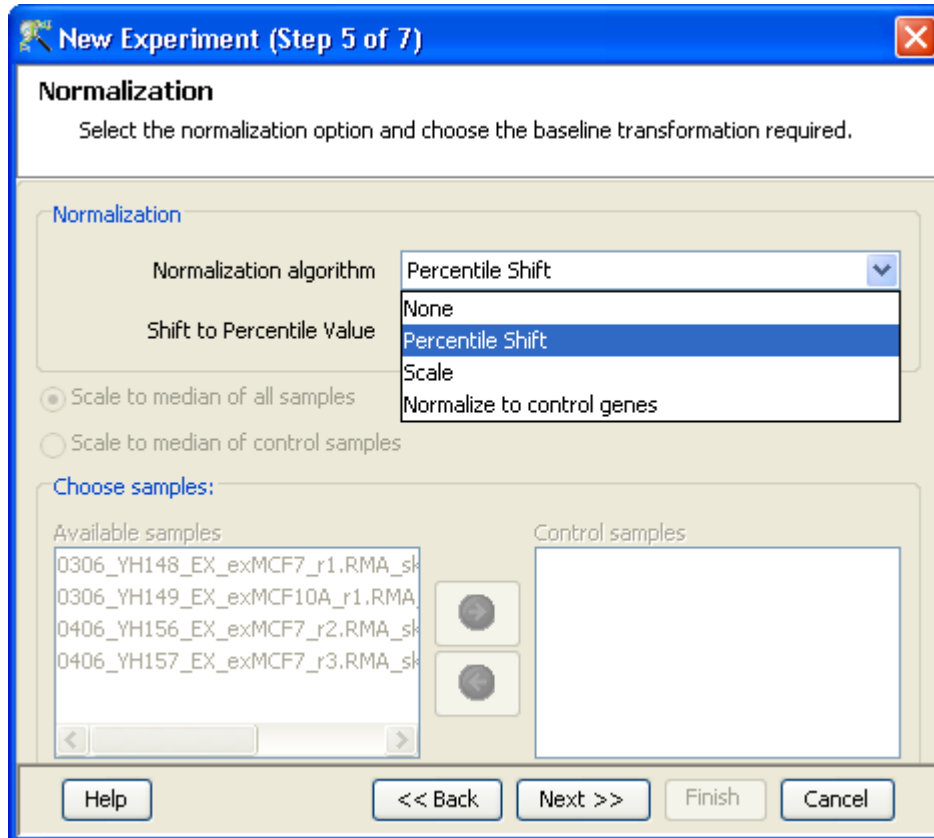


Figure 9.6: Normalization

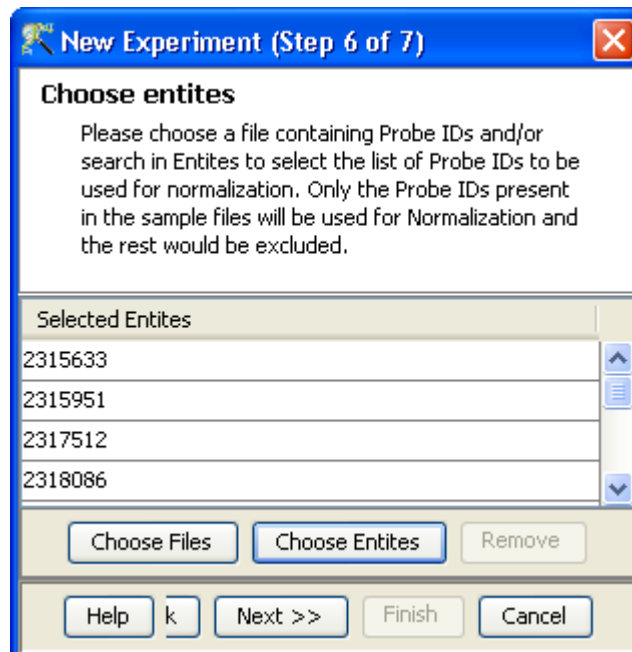


Figure 9.7: Normalize to control genes

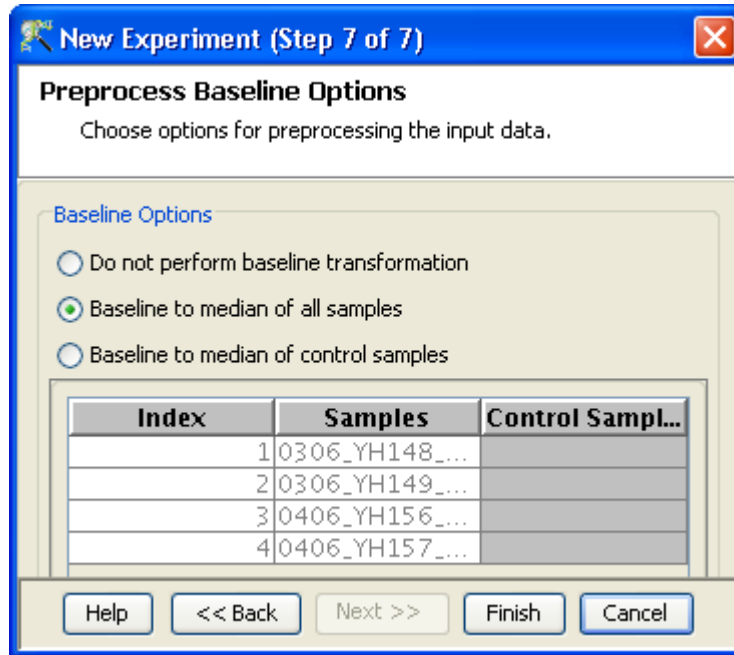


Figure 9.8: Normalization and Baseline Transformation

9.1.2 Data Processing for Exon arrays

This section describes the data processing which occurs during the experiment creation step. During the experiment creation steps, probeset level and transcript level data are processed simultaneously. Even though only the transcript level processing is user specified, the probe set level processing uses the exact same set of parameters. Thus, while for exon summarization the probes in each individual probe set are summarized, for generating transcript level data, all the probes within the transcript are summarized. DABG is then performed. DABG stands for "detection above background". It is calculated by comparing Perfect Match(PM) probes to a distribution of background probes. This comparison is used to generate a p-value. This is then combined into a probe set level p-value using the Fischer equation. This option allows the user to filter the transcripts(genes) having low expression values by correlating it with the probeset(exon) level data.

1. **File formats:** The data file should be present either as a CEL file or a CHP file. However while creating an experiment; only one type of file (CEL/CHP) can be used.
2. **Raw signal values (CEL files):** In an Affymetrix Exon Expression experiment, the term "raw" signal values refers to the linear data which has been summarized using a summarization algorithm (RMA16, PLIER 16 and Iterative PLIER 16). All summarization algorithms also do variance stabilization by adding 16. This is applicable to both the transcript and the probeset level data. Raw values for both are shown in the entity inspector.
3. **Raw signal values (CHP files):** In an Affymetrix Exon Expression experiment, the term "raw" files refers to the linear data obtained from the CHP files. This is applicable to both the transcript and the probeset level data. Raw values for both are shown in the entity inspector.

4. **Normalized signal values (CEL files):** "Normalized" values are generated after the log transformation and baseline transformation step. This is applicable to both the transcript and the probeset level data and the same transforms are performed on both. The normalized signal value of the probeset can be viewed under the signal value tab of the splicing visualization link.
5. **Normalized signal values (CHP files):** The term "Normalized" refers to values generated after log transformation, normalization (Percentile Shift, Scale or Normalize to control genes) and baseline transformation. This is applicable to both the transcript and the probeset level data. The normalized signal value of the probe set can be viewed under the signal value tab of the splicing visualization link.
6. **Gene-level Normalized intensity:** It is the difference of normalized exon-level signal and its normalized gene-level signal
7. **Treatment of on-chip replicates:** Not Applicable.
8. **Flag values:** Not Applicable.
9. **Treatment of Control probes:** Not Applicable.
10. **Empty Cells:** Not Applicable.
11. **Sequence of events (CEL files):** The sequence of events involved in the processing of a CEL file is: Summarization—→Log Transformation—→Baseline Transformation. This is applicable to both the transcript and the probeset level data.
12. **Sequence of events (CHP files):** If the data in the CHP file is already log transformed, then **GeneSpring GX** detects it and proceeds with the normalization step. This is applicable to both the transcript and the probeset level data.

9.1.3 Experiment setup

- **Quick Start Guide**

Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the *Advanced Workflow*, the method of analysis, the details of the algorithms used and the interpretation of results.

- **Experiment Grouping:** *Experiment parameters* defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different

probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.

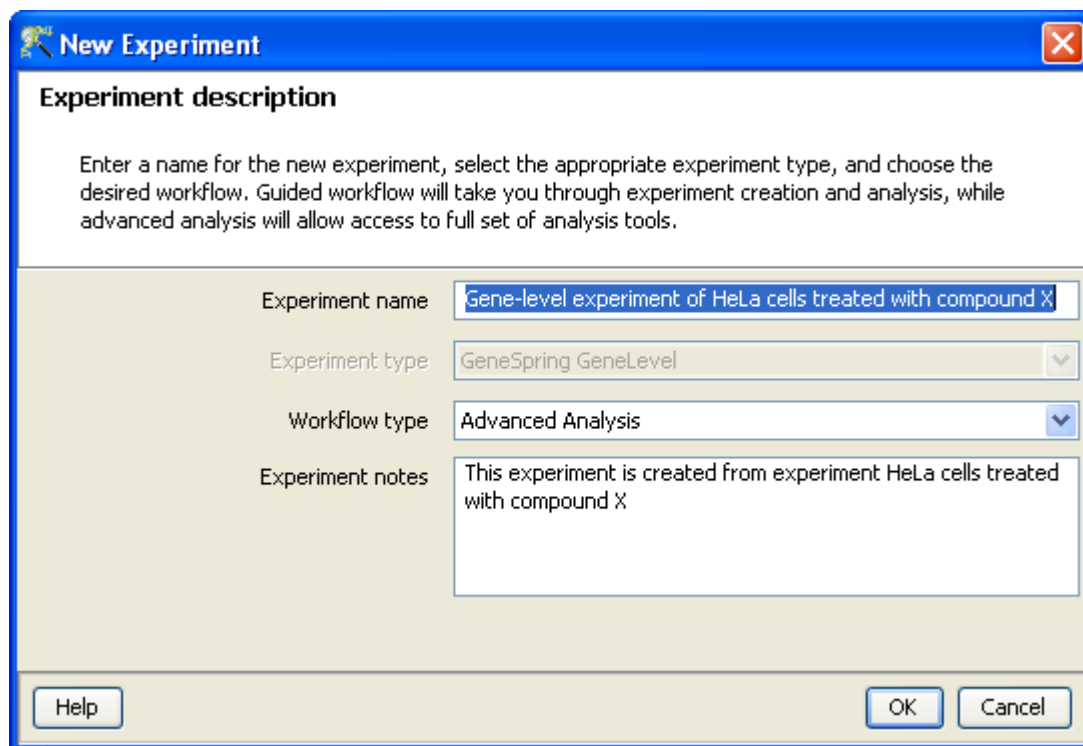


Figure 9.9: Gene Level Experiment Creation

- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

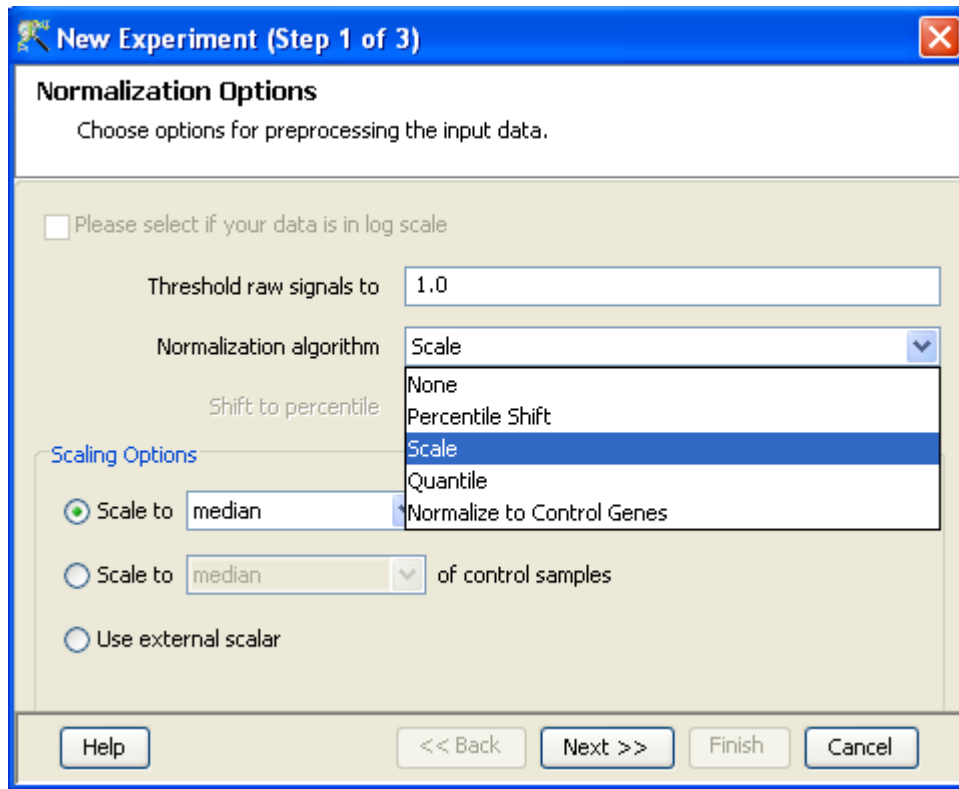


Figure 9.10: Gene Level Experiment Creation - Normalization Options

Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

Note: All links in the Workflow Browser work on transcript level data; the only exception are links in the section on Splicing Analysis. All entity lists store only transcript level data as well.

9.1.4 Quality Control

- **Quality Control on Samples**

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows three tiled windows:

- 3D PCA scores, Correlation coefficients and Correlation plot tabs.
- Experiment grouping and Hybridization Controls(available for CEL files).
- Legend.

Figure 9.13 has the 4 tiled windows which reflect the QC on samples.

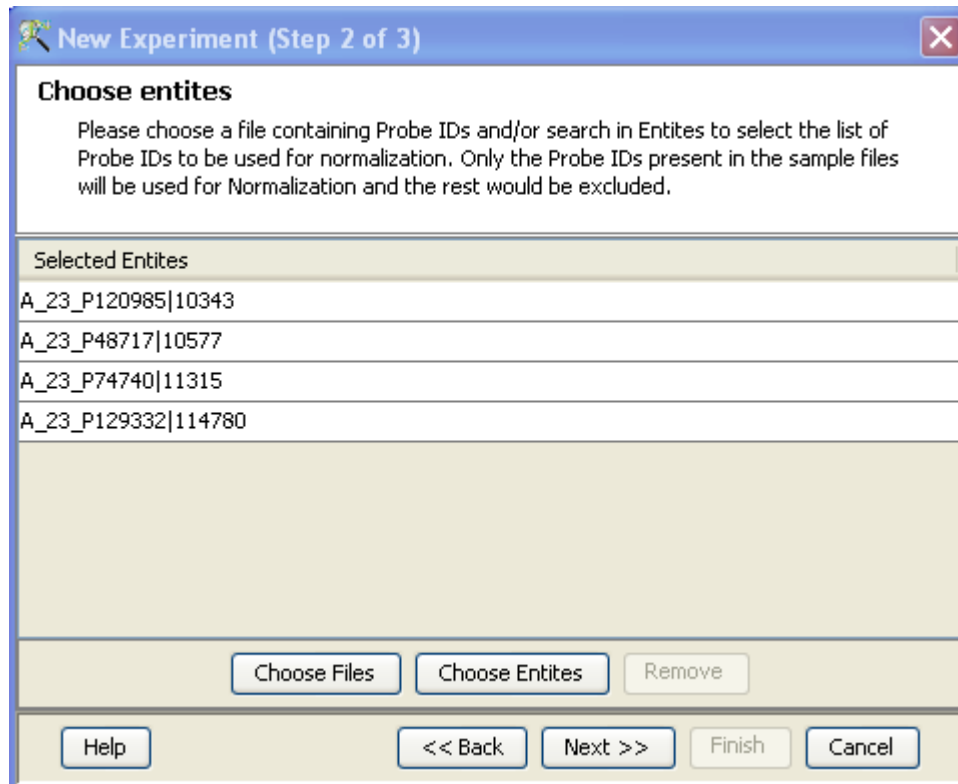


Figure 9.11: Gene Level Experiment Creation - Choose Entities

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$

The heatmap is colorable by Experiment Factor information via Right-Click→Properties. The intensity levels in the heatmap can also be customized here.

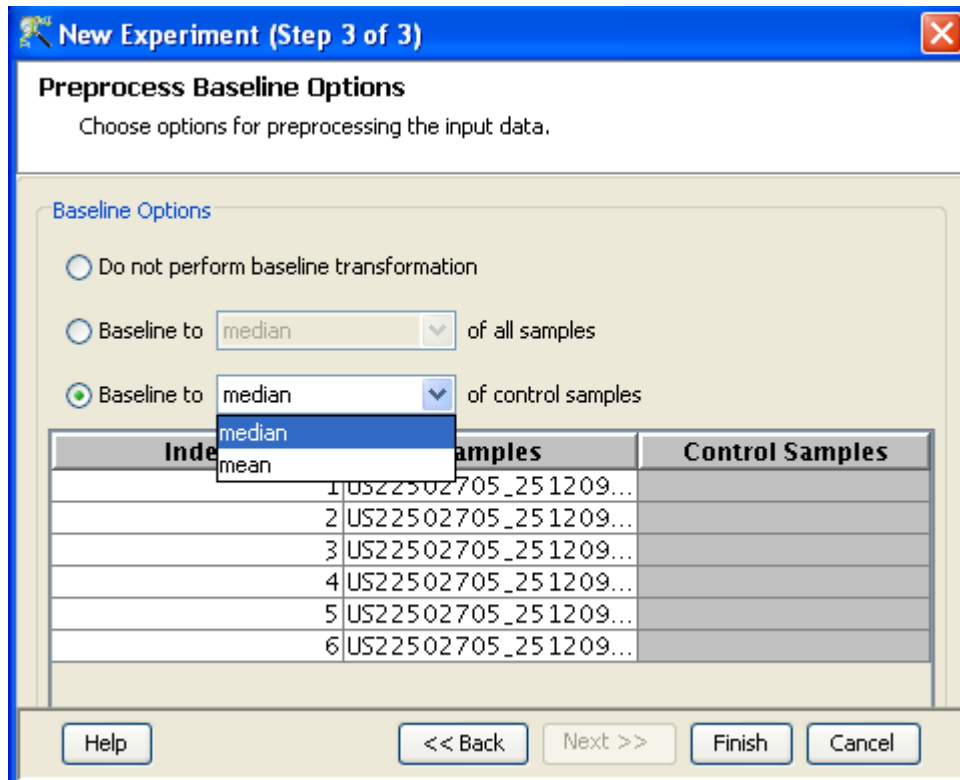


Figure 9.12: Gene Level Experiment Creation - Preprocess Baseline Options

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

Experiment Grouping shows the parameters and parameter values for each sample.

The *Hybridization Controls* view depicts the hybridization quality. Hybridization controls are composed of a mixture of biotin-labelled cRNA transcripts of bioB, bioC, bioD, and cre prepared in staggered concentrations (1.5, 5, 25, and 100pm respectively). This mixture is spiked-in into the hybridization cocktail. bioB is at the level of assay sensitivity and should be called Present at least 50% of the time. bioC, bioD and cre must be Present all of the time and must appear in increasing concentrations. The X-axis in this graph represents the controls and the Y-axis, the log of the Normalized Signal Values.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-summarization of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on *OK* to proceed.

- **Filter Probe Set by Expression** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags** No flags are generated during creation of exon splicing experiment.

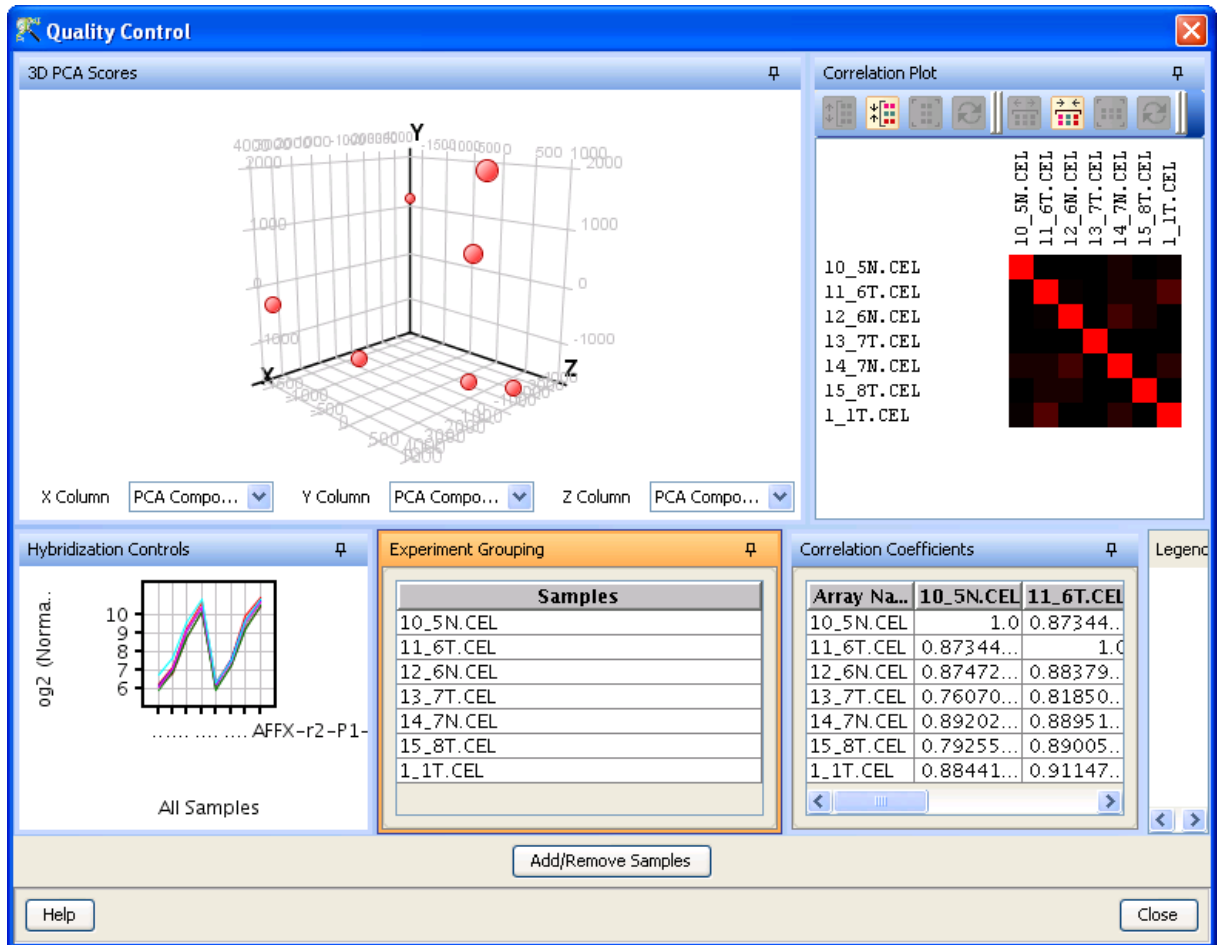


Figure 9.13: Quality Control

- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

9.1.5 Analysis

- **Statistical Analysis**
For details refer to section [Statistical Analysis](#) in the advanced workflow.
- **Filter on Volcano Plot**
For details refer to section [Filter on Volcano Plot](#)
- **Fold Change**
For details refer to section [Fold Change](#)
- **Clustering**
For details refer to section [Clustering](#)

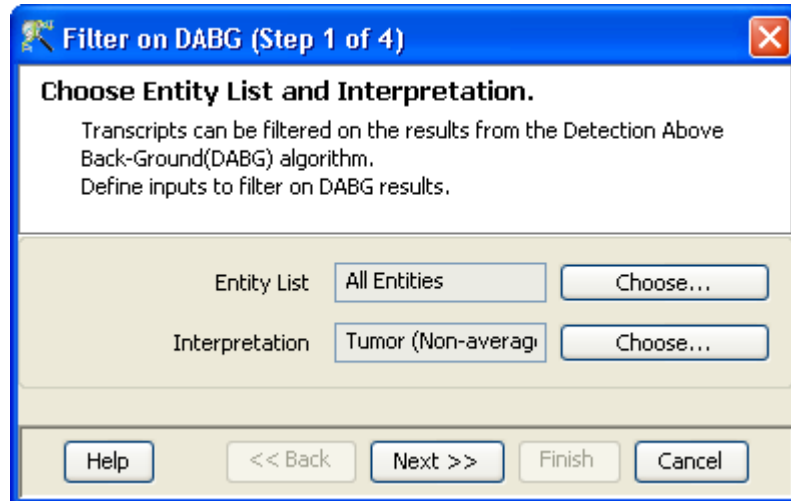


Figure 9.14: Input Data

- **Find Similar Entities**

For details refer to section [Find Similar Entities](#)

- **Filter on Parameters**

For details refer to section [Filter on Parameters](#)

- **Principal Component Analysis**

For details refer to section [PCA](#)

9.1.6 Exon Splicing Analysis

This analysis section is specific to the Affymetrix Exon Splicing Workflow. The following options are provided:

- **Filter transcripts on DABG:**

DABG is performed at the time of experiment creation. For Exon Splicing analysis, the transcripts can be filtered on DABG results. This occurs through a four step wizard and the filtering considers only core probesets (for the purpose of calling a transcript as Present) even though DABG values were generated initially for all the probe sets. The DABG values are stored only for probesets that are a part of exon summarization or if they are marked core. This change does not affect the downstream analysis and only reduces the memory usage.

1. The first step allows the user to choose the entity list and interpretation. See figure [9.14](#).
2. In the second step, the filtering options can be specified. Probe sets are defined as Present based upon a p-value cut-off which was generated during executing of DABG algorithm. Secondly, the minimum percentage of **core** exons that should be present in a gene in a sample to mark

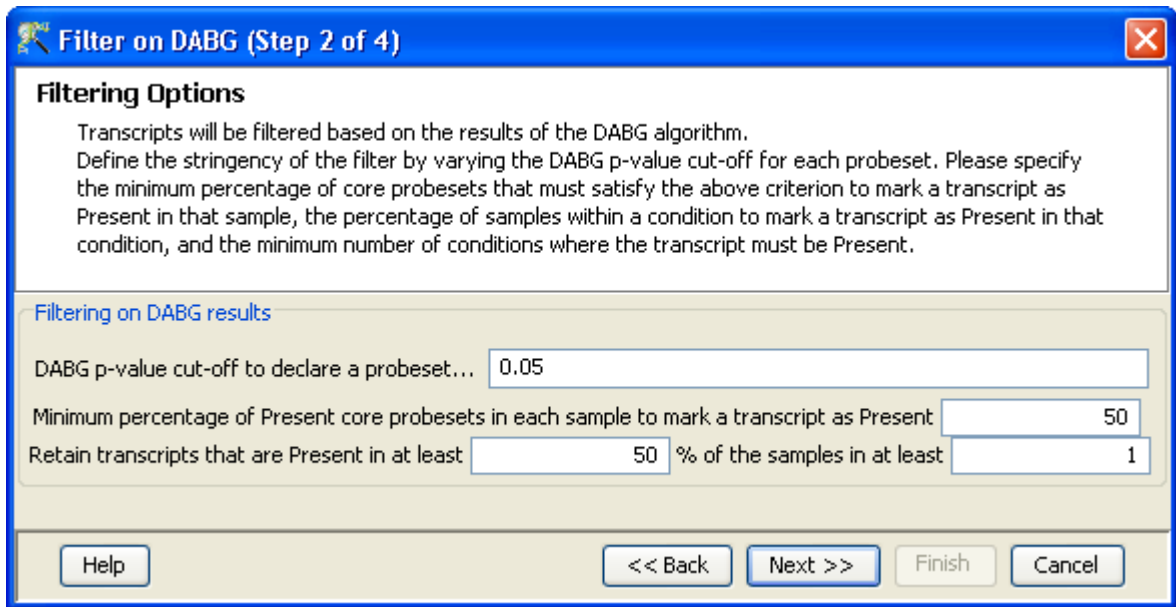


Figure 9.15: Filtering Options

it as Present should be given. For a transcript to be called as Present, a substantial number of core probe sets should be "Present" (as designated by the DABG generated p-value). The default value specifies 50% of core probe sets to be 'Present'. The percentage of samples (within a condition) in which a gene must be present for it to be retained is set at 50% and can be increased for more stringency. See figure 9.15.

3. This step shows the entities which have passed the filter, in the form of a spreadsheet (along with their normalized values) and a profile plot. The number of entities passing the filter is mentioned at the top of the panel. See figure 9.16
4. The last step shows all the entities passing the filter along with their annotations. It also shows the details (regarding creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to entities which satisfied the cutoff. Double clicking on an entity in the *Entities* table opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using *Configure Columns* button. See figure 9.17

For more details on DABG and on the defaults used in the filtering option, refer to Affymetrix white paper [3].

- **Splicing ANOVA:**

Splicing ANOVA initially calculates the gene-level normalized intensities for each of the probesets (i.e., the difference between probeset level signal and transcript level signal). Then it runs a $(n + 1)$ -way ANOVA where n denotes the number of parameters in the chosen interpretation and the plus 1 is on account of the added probeset parameter. Currently, **GeneSpring GX** supports values of only 1 or 2 for n . The alternative splicing p-value is given by the p-value for the probeset*parameter term when there is only one parameter. In the event that there two parameters, individual p-values are output for

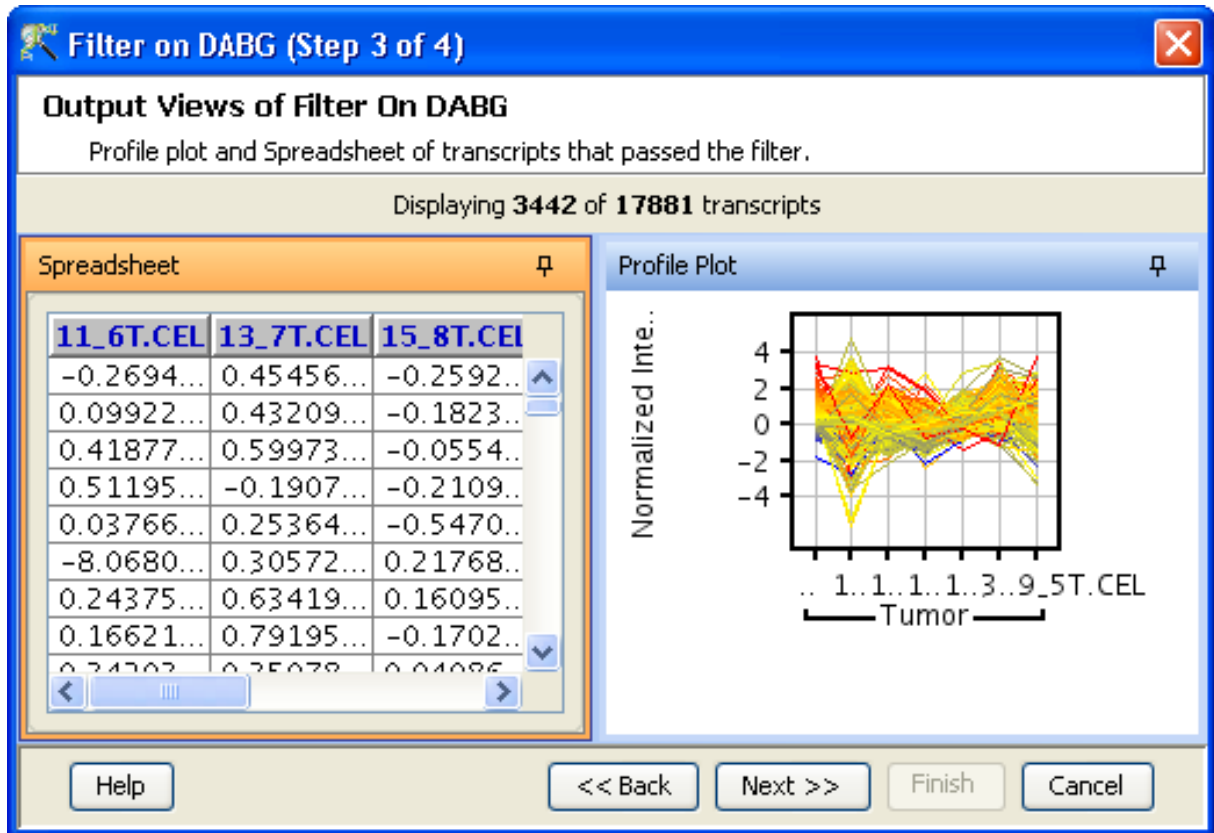


Figure 9.16: Output Views

each of $\text{probeset} * \text{parameter1}$ and $\text{probeset} * \text{parameter2}$ as well as $\text{probeset} * \text{parameter1} * \text{parameter2}$. Additionally, in situations where parameter1 and parameter2 are such that each replicate group has only one sample, only the $\text{probeset} * \text{parameter1}$ and $\text{probeset} * \text{parameter2}$ p-values are output. This could happen in paired experiments, i.e., if you have paired tumor and normal samples and 2 parameters, one indicating disease-state and other indicating the individual from whom the sample is derived; in such cases, the p-value of interest is $\text{probeset} * \text{disease-state}$. Note that both balanced and unbalanced designs are supported but balanced designs will run faster. Unbalanced designs will progress slowly for transcripts with many probesets and canceling will cause display of results on all transcripts which have completed so far.

Splicing ANOVA can be executed in 5 steps:

- This step requires the user to provide the entity list and the interpretation. See figure 9.18.
- Filtering criteria specified are provided here. See figure 9.19.
 1. This specifies the probe set list on which Splicing ANOVA should be calculated. This option is dependant on the probe set list initially used for summarization. For example, if the full list was selected then both the core and extended lists are available and if extended was selected, then core would also appear in the list of options.
 2. Filtering of probe sets is based upon the results of DABG algorithm. For more details refer to the section on [Filter transcripts on DABG](#). Note that unlike the filter transcripts steps, the goal here is to identify which probesets for a transcript should be carried into splicing ANOVA.

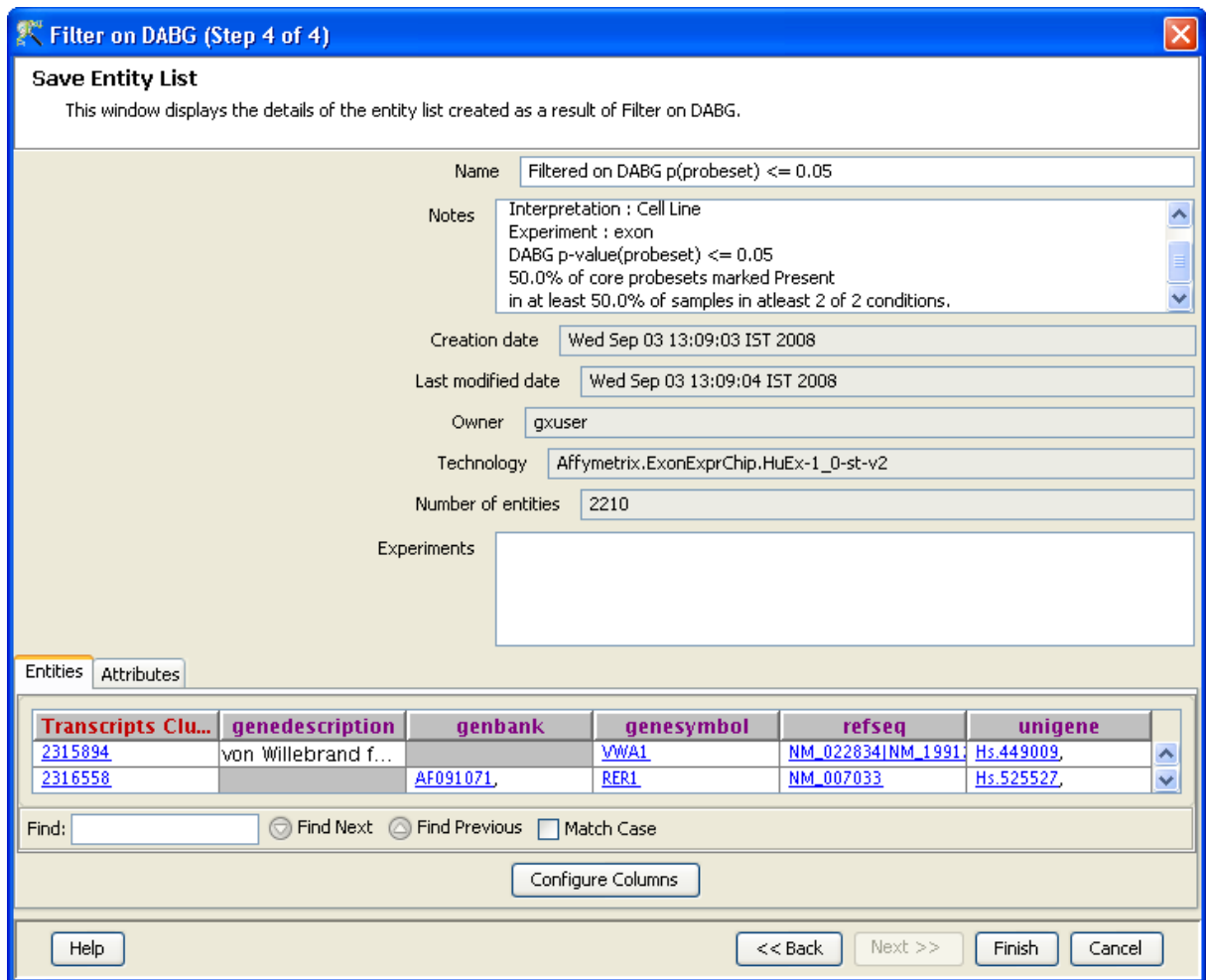


Figure 9.17: Save Entity List

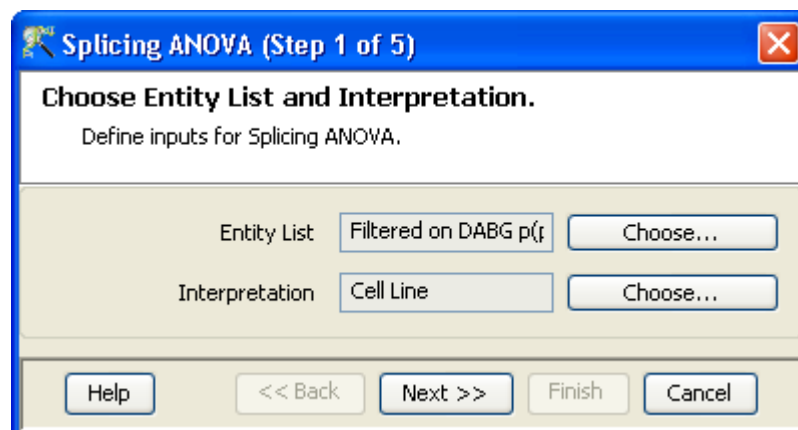


Figure 9.18: Input Data

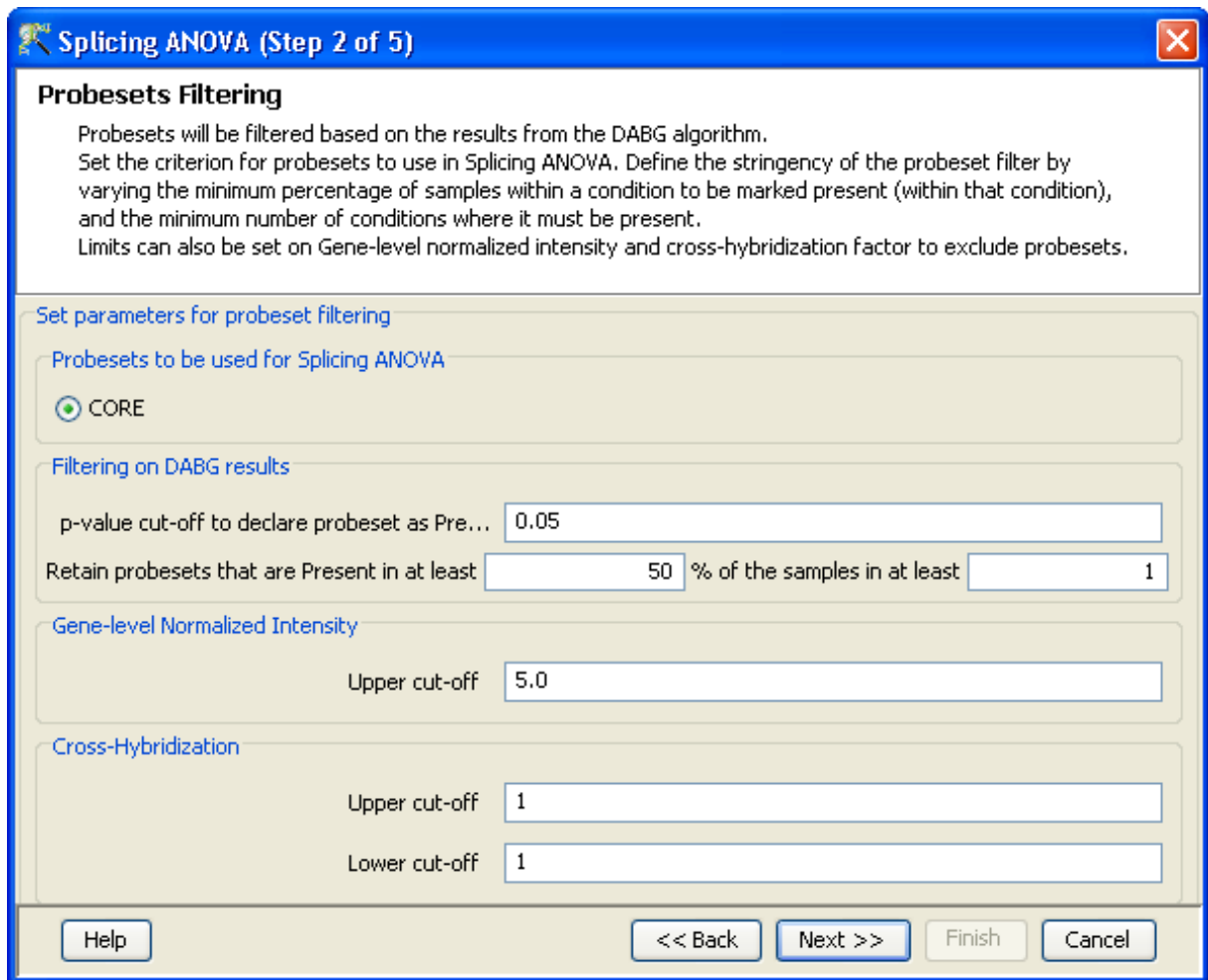


Figure 9.19: Filtering of Probesets

3. Gene-level Normalized Intensity: Probesets with large gene-level normalized intensities are excluded from ANOVA. The default is set at 5.0, which means that if the gene level normalized intensity of any probeset is greater than $\log_2(5.0)$ in a minimum of one sample, then that probe set will be excluded from splicing ANOVA. This filter is implemented to weed out probes with high background and cross hybridization potential.
 4. Cross-hybridization: Probe sets with high cross-hybridization potential are removed from the analysis. Only probesets with value of 1 have been recommended to be included in analysis. Refer to [3] for more details.
- The multiple testing correction to be implemented for p-value computation is chosen here. See figure 9.20
 - This step shows the results of the Splicing ANOVA in the form of a spreadsheet. For each transcript, the p-value, corrected p-value and the number of probesets that were for performing Splicing ANOVA are shown. If multiple p-value are computed, then the list of transcripts shown are exactly those for which any one of the p-values is within the specified threshold. The default p-value cut-off used is 0.05 but it can be reconfigured using the *Change p-value cut-off* button. Transcripts in which only one probeset has passed the previously applied filters are automatically excluded. The term PROBESET in the p-value names indicates that the p-value

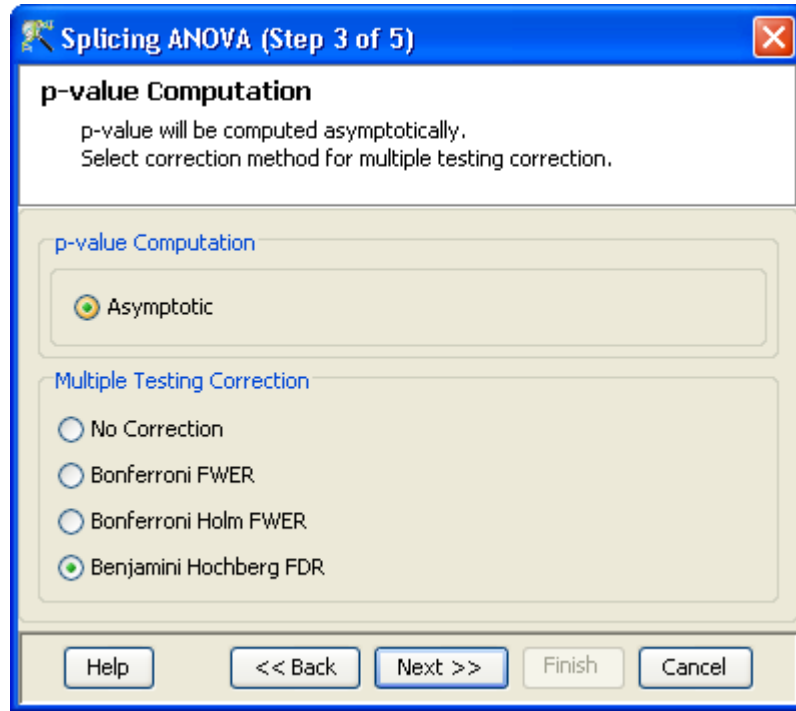


Figure 9.20: Multiple Testing Correction

is for an interaction term with the probeset parameter. See figure 9.21.

- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using **Configure Columns** button. Selecting **Finish** results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator. The Entity List generated as a result of Splicing ANOVA has an attachment associated with it. The attachment remembers which probesets were used to perform splicing ANOVA for each transcript. The 'Splicing Visualization' step will use this attachment to show relevance probesets for a transcript in the variance plots. This attachment is also carried over when a custom list is created while performing 'Filter on Splicing Index' or during the 'Splicing Visualization' steps. See figure 9.22.

For more details on Splicing ANOVA and the defaults specified in this option, refer to Affymetrix white papers [5, 3]

- **Filter on Splicing Index:**

Splicing Index is essentially a fold change analysis step wherein difference between the gene normalized signal intensities for 2 conditions are computed as follows:

- For a given transcript, this difference is computed for each probeset; if any of the probesets has an absolute value difference greater than the specified threshold (0.5 by default) then the transcript will pass this filter.
- In situations where the interpretation has only 1 condition, the Splicing Index is computed against zero.

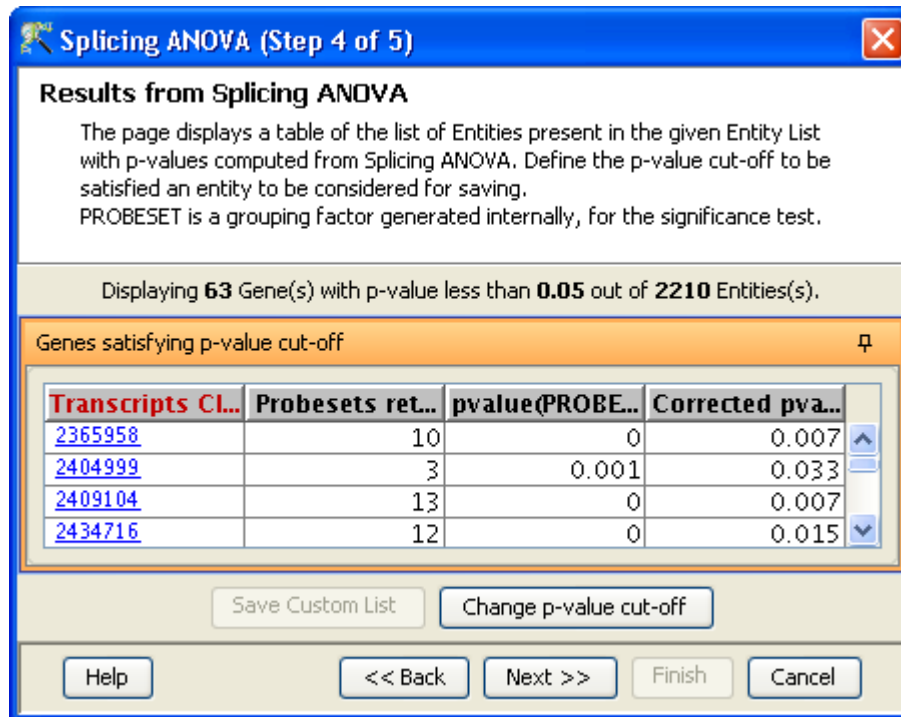


Figure 9.21: Results

This analysis is executed in four steps:

1. This step requires the user to provide the entity list and the interpretation. See figure 9.23.
2. The second step in the wizard asks the user to select pairing options based on parameters and conditions in the selected interpretation. In case of two or more groups, the user can evaluate either pairwise or with respect to a control. In the latter situation, the condition to be used as control needs to be specified. The order of conditions can also be flipped (in case of pairwise conditions) using an icon. See figure 9.24
3. This step shows the results of the analysis in the form of a spreadsheet. The transcripts that have passed the cut-off are shown along with the Splicing Index. It also displays the probesets considered for each transcript, for calculating the Splicing Index. The cut-off can be changed using the *Change Splicing Index cut-off button*. See figure 9.25
4. The last step shows all the entities passing the filter along with their annotations. It also shows the details (regarding creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to entities which satisfied the cutoff. Double clicking on an entity in the **Entities table** opens up an **Entity Inspector** giving the annotations corresponding to the selected profile. Additional tabs in the **Entity Inspector** give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using *Configure Columns* button. See figure 9.26

- **Splicing Visualizations:**

The results of splicing analysis can be viewed as 6 tabs under the 'Splicing Visualization' link in the workflow.

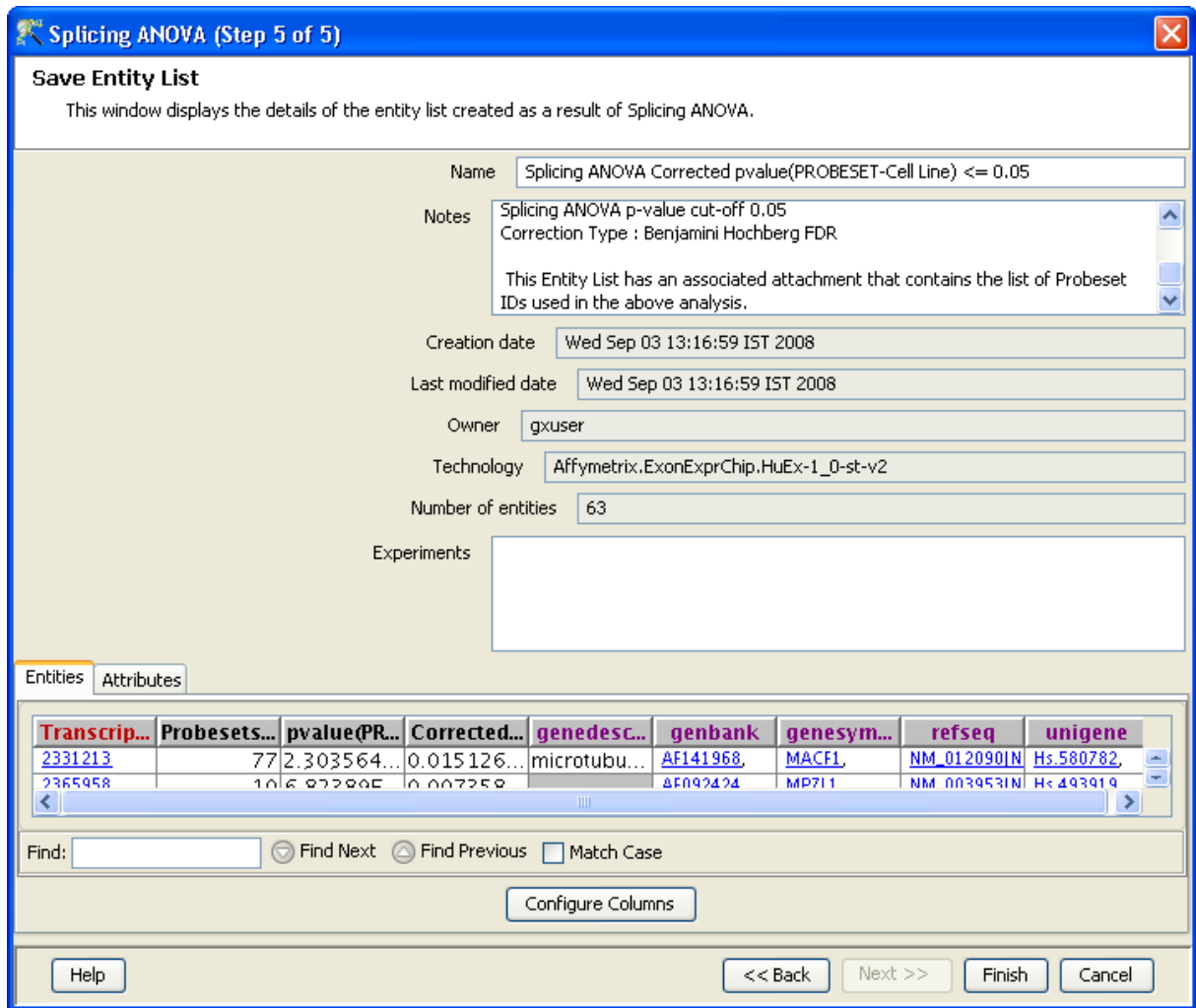


Figure 9.22: Save Entity List

- This step requires the user to provide the entity list and the interpretation. See figure 9.27.
- The next step presents the data in 6 views. The first three show normalized probeset signals while the last three show gene normalized probeset signals. The legend for the current view is present on the left and a message on the top shows the gene being displayed. The tabs for the views are present at the bottom along with the option of selecting the transcript(gene) to be viewed. Only one transcript can be viewed at a time. Clicking on **Save Transcript** adds the transcript in view to a cache, which is then saved as a new entity list when the wizard concludes. See figure 9.28.
 - * **Signal Values:** This displays the normalized intensity values of probesets in the selected transcript. The exons corresponding to the probesets as well as associated annotation information on the probesets such as chromosomal location and level are also given. In case the entity list used is obtained after *Splicing ANOVA*, a column containing information on whether the probeset was filtered out or used for splicing ANOVA is also given.
 - * **Probeset Profile Plot:** This shows a profile plot of the probesets in the selected transcript. When run on an output list from *Splicing ANOVA* or *Splicing Index*, the grayed out profiles, if any, indicate probesets filtered out in *Splicing ANOVA*. The data used in the plot is the

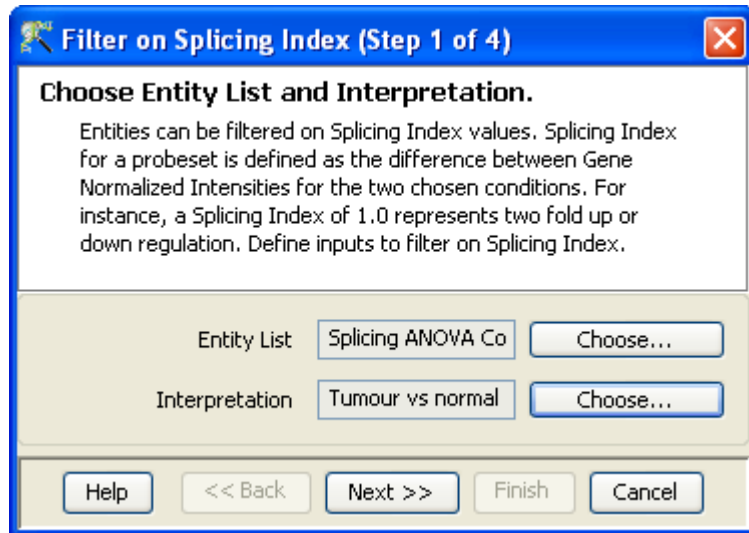


Figure 9.23: Input Data

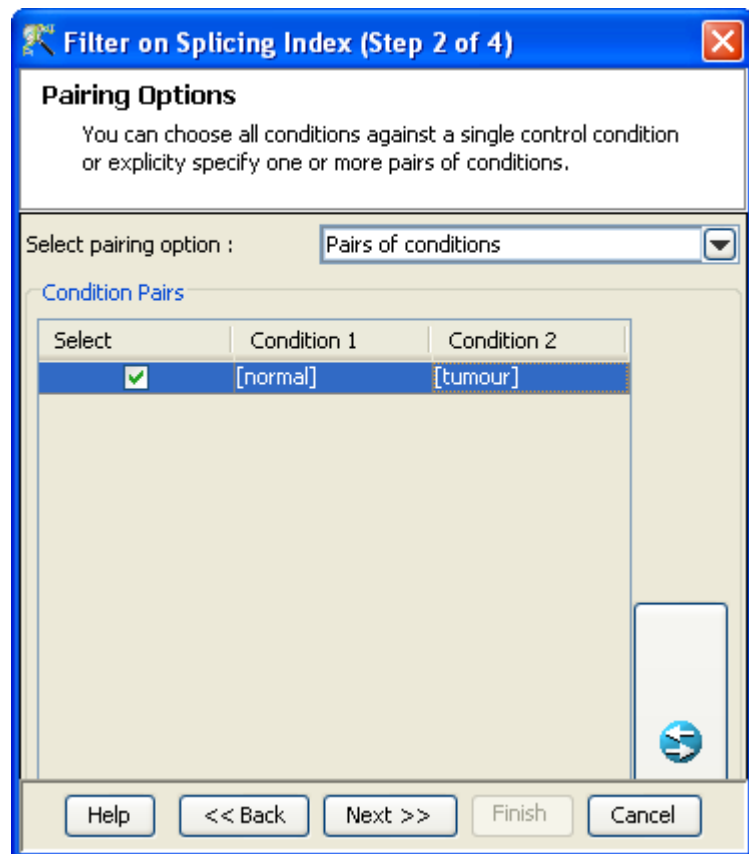


Figure 9.24: Pairing Options

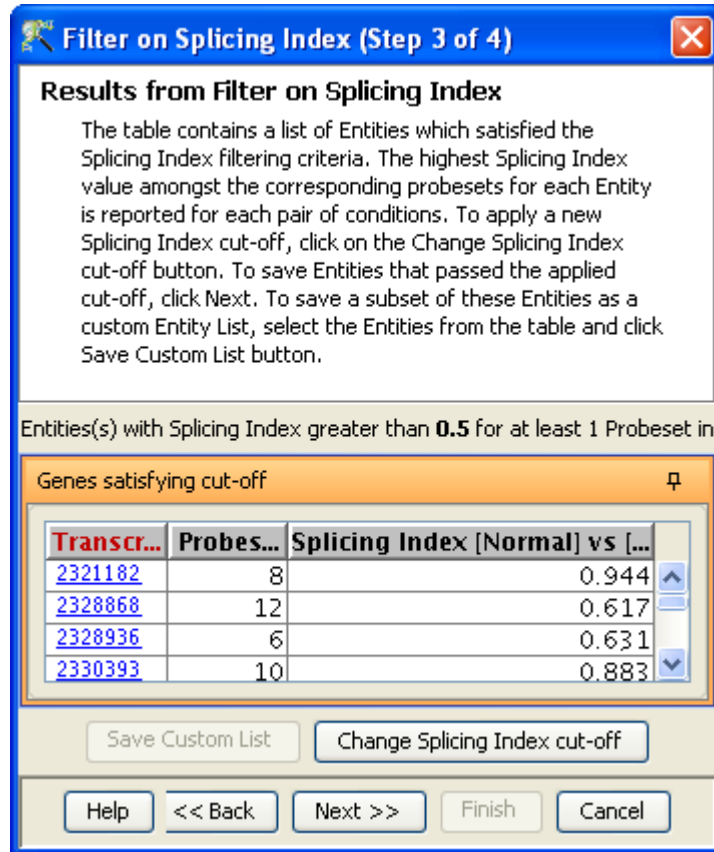


Figure 9.25: Results

probeset level normalized data.

- * **Probeset Variance Plot:** This shows the variance in the probesets across different conditions. The X-axis contains the probeset ID ordered by chromosomal location and the Y-axis is the mean of the probeset normalized intensity values across samples in a condition. The plot shows each point on the profile as a shape (where the shape determines the exon as described in the legend). The plot also shows error bars for each of the points in the profile, where the error bars indicate the standard error of mean within the corresponding condition (the standard error of mean is defined as the standard deviation divided by the square root of number of samples in the condition minus 1)
- * **Gene Normalized Signal Values:** This is similar to the *Signal Values* view except that the intensities shown are 'Gene Normalized Signals'. The gene normalized signal refers to the difference between the probeset level normalized signal and the transcript level normalized signal.
- * **Gene Normalized Profile Plot:** This shows a profile plot of the probesets in the selected transcript. The greyed out profiles belong to the exons filtered out in *Splicing ANOVA*. The data used in the plot is the gene normalized data.
- * **Gene Normalized Variance Plot:** This is similar to the *Probeset Variance Plot* except that the intensity values used are the gene-level normalized intensities. This is often the most useful plot for viewing splicing and therefore also the default view.

In case of various probeset IDs corresponding to the same exon in a transcript, they usually have



Filter on Splicing Index (Step 4 of 4)



Save Entity List

This window displays the details of the entity list created as a result of Filter on Splicing Index.

Name

Notes

Creation date

Last modified date

Owner

Technology

Number of entities

Experiments

Entities

Attributes

Transcrip...	Probesets...	Splicing I...	transcript...	cat
2375706	27	0.9356602	2375706	mair

Find: Match

253

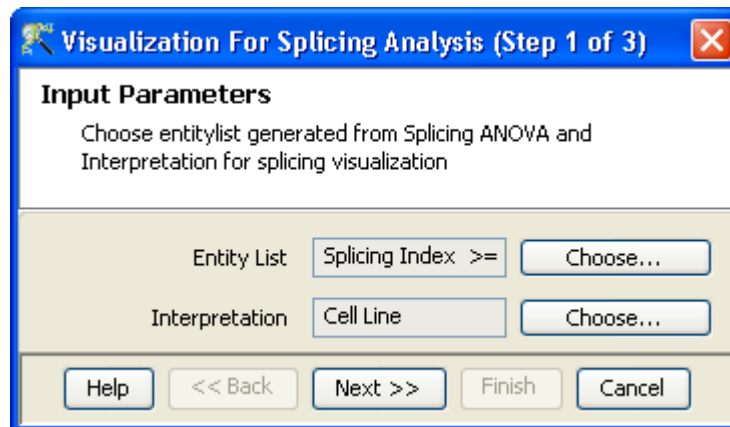


Figure 9.27: Input Data

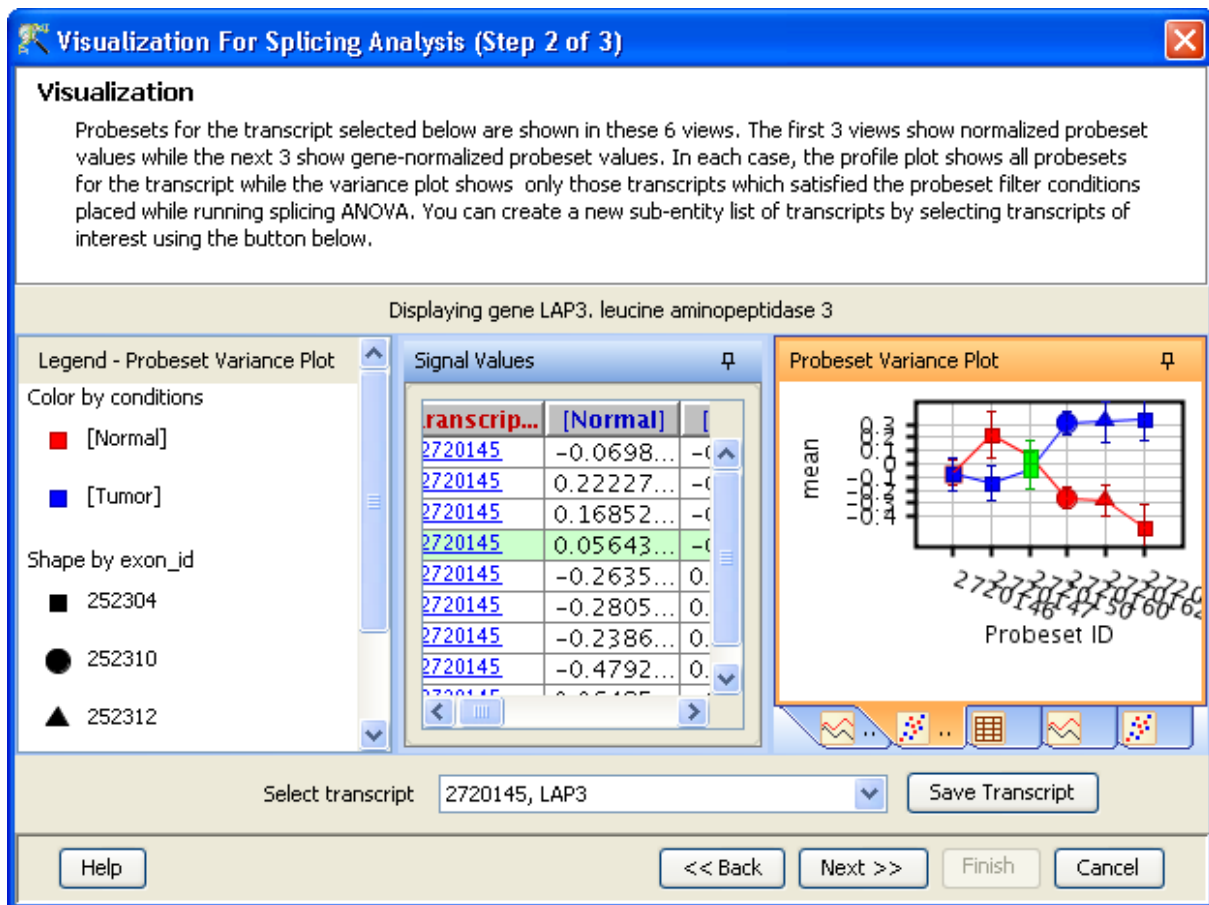


Figure 9.28: Visualization

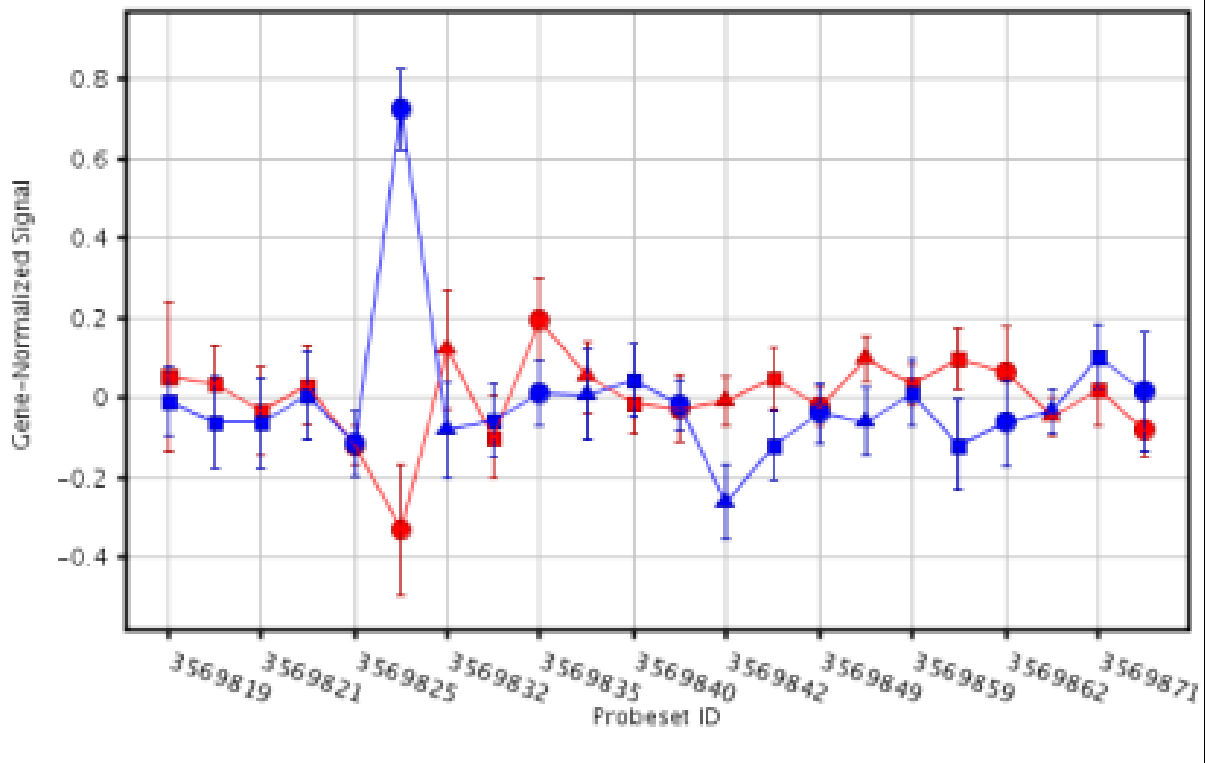


Figure 9.29: Visualization

similar values across a given condition. In case there is a significant difference in the expression levels for a particular probeset when compared to others (all of them having the same exon ID), then it could either mean:

1. The expression level of that particular probeset might be erroneous owing to noise or other experimental factors.
 2. The particular probeset might actually correspond to a sub-exon within the exon.
- The third step appears only when a transcript is saved. The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using *Configure Columns* button. Selecting *Finish* results in an entity list being created containing selected entities. The name of the entity list will be displayed in the experiment navigator. See figure 9.30.

9.1.7 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

Visualization For Splicing Analysis (Step 3 of 3)

Save Entity List

This window displays the details of the entity lists created as chosen in last page.

Name:

Notes:

Creation date:

Last modified date:

Owner:

Technology:

Number of entities:

Experiments:

Entities

Transcrip...	genedesc...	genbank	genesym...	refseq	unigene
2720145	leucine a...	AK130293	LAP3	NM_015907	Hs.570791,

Find: Match Case

Figure 9.30: Save Entity List

9.1.8 Results

- **Gene Ontology (GO) analysis**

GO is discussed in a separate chapter called [Gene Ontology Analysis](#).

- **Gene Set Enrichment Analysis (GSEA)**

Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).

- **Gene Set Analysis (GSA)**

Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).

- **Pathway Analysis**

Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).

- **Find Similar Entity Lists**

This feature is discussed in a separate section called [Find Similar Entity Lists](#)

- **Find Significant Pathways**

This feature is discussed in a separate section called [Find Significant Pathways](#).

- **Launch IPA**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Import IPA Entity List**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Extract Interactions via NLP**

This feature is discussed in detail in the chapter [Pathway Analysis](#).

9.1.9 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)

- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)

- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)

- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

9.1.10 Algorithm Technical Details

Here are some technical details of the Exon RMA16, Exon PLIER16, and Exon IterPLIER16 algorithms.

Exon RMA 16. Exon RMA does RMA background correction followed by Quantile normalization followed by a Median Polish probe summarization, followed by a Variance Stabilization of 16. An option for GCBG background correction is available from *Tools* → *Options* → *Affymetrix Exon Summarization Algorithms* → *Exon RMA*. GCBG background correction bins background probes into 25 categories based on their GC value and corrects each PM by the median background value in its GC bin. Only antigenomic probes are used by default for GCBG calculation. RMA does not have any configurable parameters.

Exon PLIER 16. Exon PLIER does Quantile normalization followed by the PLIER summarization using the PM or the PM-GCBG options (the latter is default), followed by a Variance Stabilization of 16. The PLIER implementation and default parameters are those used in the Affymetrix Exact 1.2 package. PLIER parameters can be configured from *Tools* → *Options* → *Affymetrix Exon Summarization Algorithms* → *Exon PLIER/IterPLIER*.

Exon IterPLIER 16. Exon IterPLIER does Quantile normalization followed by the IterPLIER summarization using the PM or the PM-GCBG options (the latter is default), followed by a Variance Stabilization of 16. IterPLIER runs PLIER multiple times, each time with a smaller subset of the probes obtained by removing outliers from the previous PLIER run. IterPLIER parameters can be configured from *Tools* → *Options* → *Affymetrix Exon Summarization Algorithms* → *Exon PLIER/IterPLIER*.

Note:

- By default, only anti-genomic probes are used for background correction for RMA 16, PLIER 16 and IterPLIER 16. This can be changed by the user by going to *Tools* → *Options* → *Affymetrix Exon Summarization Algorithms*. The choice made for background probes here is applicable for the DABG p-value calculation as well.
- When RMA 16 is chosen as the transcript level summarization algorithm, the same algorithm will also be used for exon-level summarization. If PLIER 16 or IterPLIER 16 is chosen for transcript level summarization, then PLIER 16 is used for exon-level summarization.

9.2 Tutorial for Exon Splicing Analysis

GeneSpring GX provides a unique analysis tool for analyzing Affymetrix exon chip to study exon splicing. The following tutorial describes the steps in Exon Splicing Analysis using the tool. The dataset used in the tutorial can be downloaded from http://www.affymetrix.com/support/technical/sample_data/

`exon_array_data.affx`. Using the tutorial, the splicing events described in the paper "Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array" by Turpaz *et al.*, 2006 can be observed.

The tutorial will not go into the details of the experiment creation as it has been described in detail above.

1. Experiment Creation:

- Create a new experiment with experiment type 'Affymetrix Exon Splicing'
- Choose 'PLIER 16' as the Summarization algorithm in Step 4.
- Use default parameters in all other steps.
- Click *Finish*.

Experiment creation will now commence. Experiment creation includes background correction of all probe sets using the DABG algorithm.

2. Experiment Grouping:

- Group your data into 2 groups, Normal and Tumor, using 'Experiment Grouping' in the Workflow. All files whose filename ends with '_N.cel' are healthy tissue files, whereas the ones with the suffix '_T.cel' are tumor tissue files.
- Create an Interpretation with these 2 conditions.

3. Exon Splicing Analysis:

There are 4 stages in Exon Splicing Analysis.

- **Filter Transcripts on DABG:**

DABG algorithm is executed on all the probesets at the time of experiment creation. The p-values generated as a result of DABG is used at this step to filter out transcripts before proceeding for ANOVA.

- (a) Click on **Filter Transcripts on DABG** in the Workflow.
- (b) Choose 'All Entities' as the entity list and 'Tumor vs Normal' as the interpretation.
- (c) Use the default parameters in the Step 2.
- (d) In Step 3, the filtering results will be displayed as a Profile Plot, showing the profiles of the transcript clusters that passed the filter criteria.
- (e) Continue on to Step 4 and click *Finish*. A new entity list named 'Filtered on DABG p(probeset) ≤ 0.05 ' will appear in the Analysis folder in the Project Navigator.

- **Splicing ANOVA:**

Among the transcripts identified, the probesets can be eliminated or retained for Splicing ANOVA based upon the DABG generated p-values. Additional filtering can also be performed at this stage to eliminate probes with high background and cross-hybridization potential.

- (a) Click on **Splicing ANOVA** in the Workflow.

- (b) Choose 'Filtered on DABG $p(\text{probeset}) \leq 0.05$ ' as the entity list, and 'Tumor vs Normal' as the interpretation.
 - (c) Retain the defaults provided in steps 2 and 3. The transcripts that have a p-value ≤ 0.05 after running the Splicing ANOVA test will be shown.
 - (d) Continue through the steps to save this list as an entity list named 'Splicing ANOVA corrected pvalue(PROBESET-tumor vs normal)'.
- **Filter on Splicing Index:**
 Splicing Index is defined as the difference between the gene normalized signal intensities of the probesets for the normal and tumor samples. For each transcript, this fold change value is computed for all the probesets that have passed the splicing ANOVA.
 - (a) Click on **Filter on Splicing Index** in the Workflow.
 - (b) Choose 'Splicing ANOVA corrected pvalue(PROBESET-tumor vs normal)' as the entity list , and 'Tumor vs Normal' as the interpretation.
 - (c) Continue through the process and save the entity list named 'Splicing Index ≥ 0.5 '.
 - **Splicing Visualizations:**
 - (a) To visualize the results of this analysis, click on the final step in this section, **Splicing Visualizations**.
 - (b) Choose 'Splicing Index ≥ 0.5 ' as the entity list, and 'Tumor vs Normal' as the interpretation.
 - (c) A visualization results window containing 6 tabs, opens up. 3 of the tabs contain views of the gene normalized data, whereas the other 3 show the raw data.
 - (d) Click on the 'Gene normalized variance plot' tab. This plot will be most useful in finding exons that vary between the 2 experiment conditions (Normal and Tumor). To view plots for different transcripts, select the particular transcript cluster ID in the drop down box at the bottom of the visualization panel.

In case of colon cancer, splicing occurs in transcripts involved in cytoskeletal organization, ACTN1 being one of them. The Gene Normalized Variance Plot for the ACTN1 transcript shows a clear variance for the 2 conditions for Exon 3569830, whereas the other exons are fairly invariant between the 2 conditions (Fig. 9.31). This indicates that this exon is spliced out in one of the conditions.

Click on the tab immediately preceding this, i.e., the Gene Normalized Profile Plot. This plot shows the exons (if any) which were filtered out, in gray ((Fig. 9.32).

To conclude the exon splicing analysis, select each transcript you wish to save in a list, one by one in the 'Select transcript' drop down box, and click on **Save transcript** to save it. All transcripts thus chosen will appear in a new entity list 'transcripts with alternative splicing' in the Project Navigator.

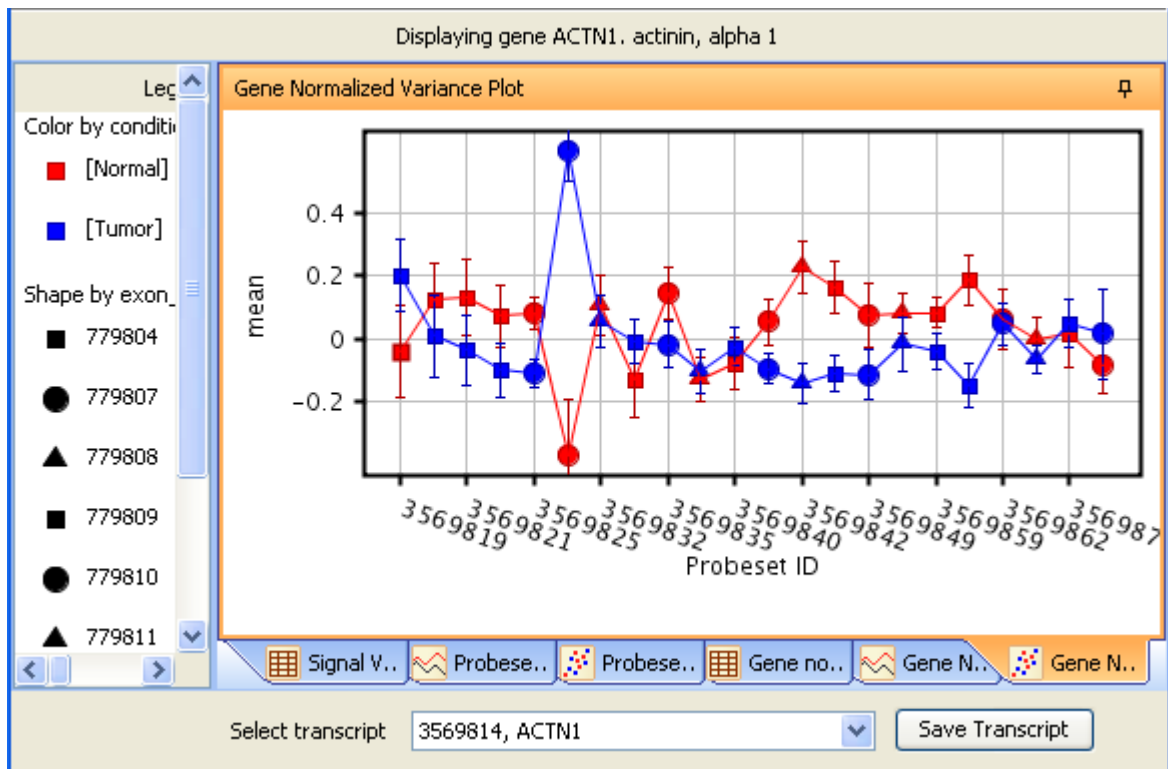


Figure 9.31: Gene Normalized Variance Plot

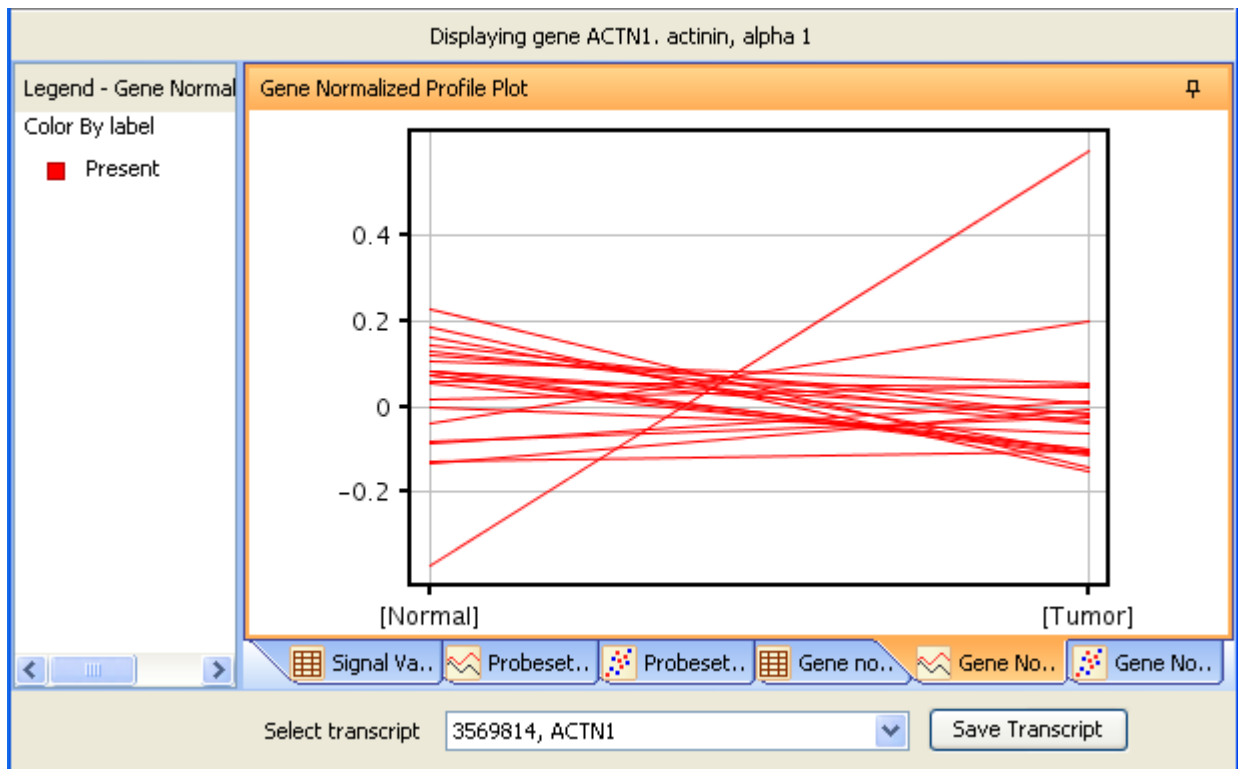


Figure 9.32: Gene Normalized Profile Plot

Chapter 10

Analyzing Illumina Data

GeneSpring GX supports the Illumina Single Color (Direct Hyb) experiments. **GeneSpring GX** supports only those projects from Genome Studio which were created using the bgx manifest files. To generate the data file, the Sample Probe Profile should be exported out from Bead Studio in **GeneSpring GX** format. These text files can then be imported into **GeneSpring GX** . From these text file, the

- Probe ID,
- Average Signal values and the
- detection p-value columns

are automatically extracted and used for project creation. Typically, a single Illumina data file contains multiple samples.

Genome Studio provides the option of performing normalization on the data, therefore if the data is already normalized, the workflow to be chosen is Advanced Analysis. This is because, *Advanced Workflow* allows the user to skip normalization steps whereas in *Guided Workflow*, normalization is performed by default.

Projects from Genome Studio created using .xml files can still be analyzed in **GeneSpring GX** , via the Custom technology creation or as Generic Single Color experiments. For more details on the same, see the section on [Illumina Custom Technology creation](#)

10.1 Running the Illumina Workflow:

Upon launching **GeneSpring GX** , the startup is displayed with 3 options.



Figure 10.1: Welcome Screen

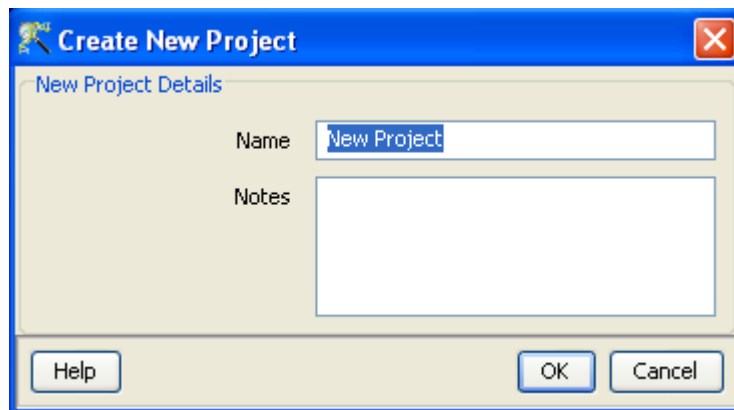


Figure 10.2: Create New project

- **Create new project**
- **Open existing project**
- **Open recent project**

Either a new project can be created or a previously generated project can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on **OK** to proceed.

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

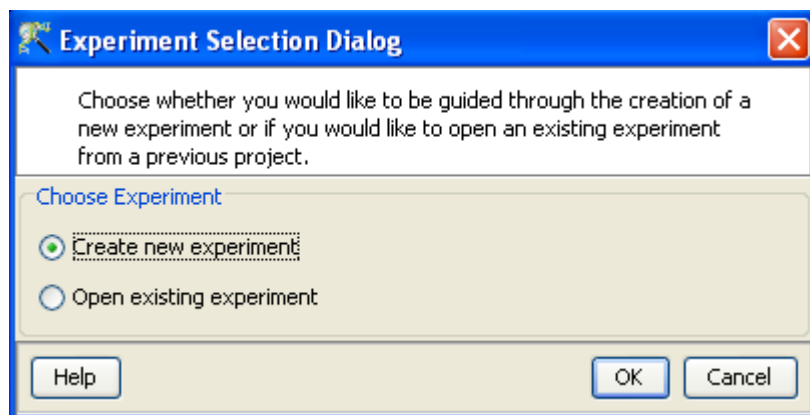


Figure 10.3: Experiment Selection

1. **Create new experiment:** This allows the user to create a new experiment. (steps described below).
2. **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the experiment type gives the user the option to choose between the multiple experiment types namely Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Agilent miRNA, Generic Single Color, Generic Two Color, Pathway and RealTime-PCR experiment.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Guided Workflow**
2. **Advanced Analysis Workflow**

Guided Workflow is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Advanced Analysis**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples

3. Reorder
4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. On the other hand, a sample is created within **GeneSpring GX**, when it associates the data files with its appropriate technology (See the section on [Technology](#)). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the *Choose Samples* option. For selecting data files and creating an experiment, click on the *Choose File(s)* button, navigate to the appropriate folder and select the files of interest. Click on **OK** to proceed.

The technology specific for any chip type needs to be created or downloaded only once. Thus, upon creating an experiment of a specific chip type for the first time, **GeneSpring GX** prompts the user to download the technology from the update server. If an experiment has been created previously with the same technology, **GeneSpring GX** then directly proceeds with experiment creation. Clicking on the *Choose Samples* button, opens a sample search wizard, with the following search conditions:

1. **Search field:** Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition:** Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on **Add** button, or can be removed from the list using *Remove* button.

Figures [10.4](#), [10.5](#), [10.6](#) show the process of choosing experiment type, loading data and choosing samples

The *Guided Workflow* wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The Workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

10.2 Data Processing for Illumina arrays

- **File formats:** The data file (.txt format) should be the Sample Probe Profile that is exported out from Bead Studio in **GeneSpring GX** format.

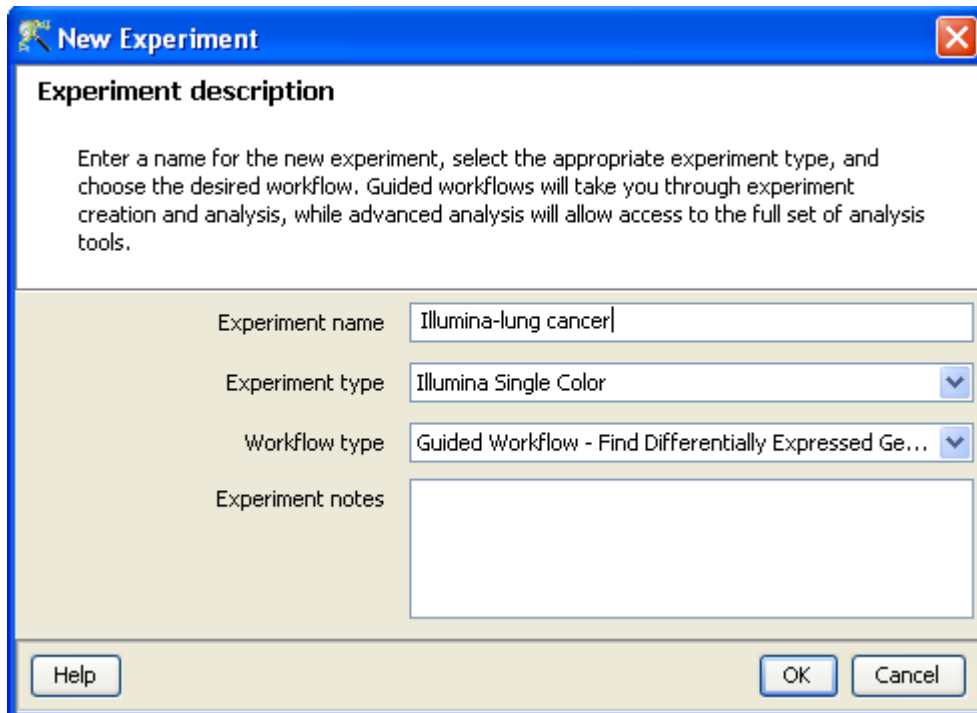


Figure 10.4: Experiment Description

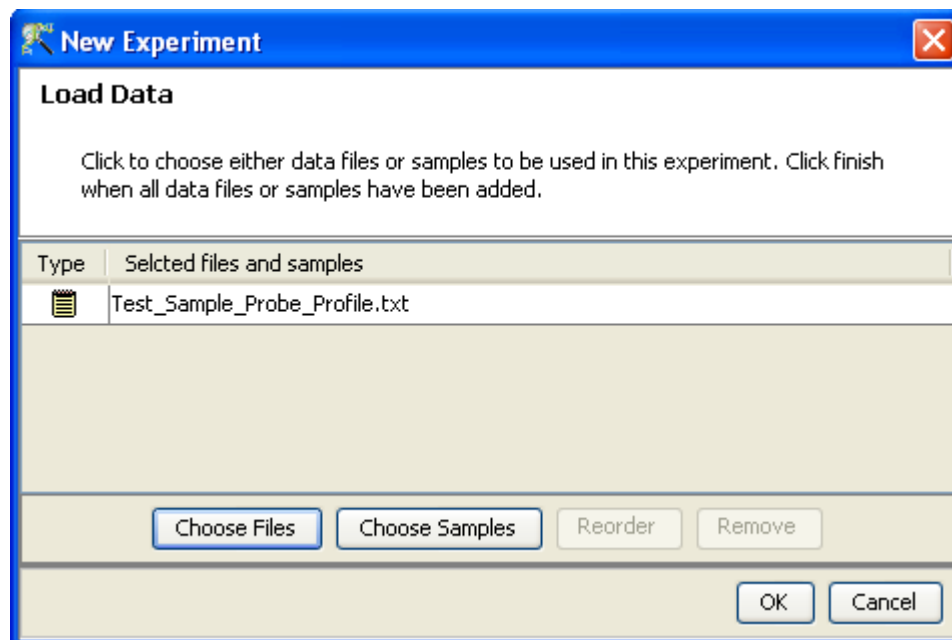


Figure 10.5: Load Data

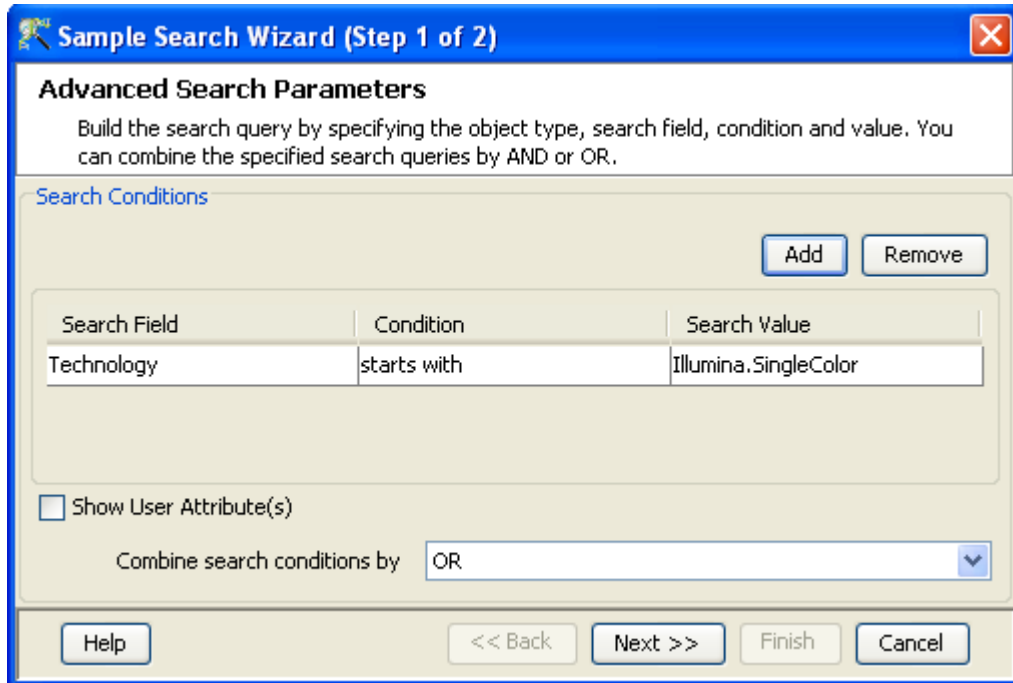


Figure 10.6: Choose Samples

- **Raw Signal Values:** The term "raw" signal values refer to the linear data that is present in the data file that is imported into **GeneSpring GX** from Genome Studio.
- **Normalized Signal Values:** "Normalized" value is the value generated after thresholding, log transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and Baseline Transformation.
- **Treatment of on-chip replicates:** It is not applicable as the data obtained from Genome Studio is already summarized.
- **Flag values:** The flag values are calculated based on the detection p-value column (from Genome Studio) and the flag settings defined by the user in the second step of experiment creation in the Advanced Workflow. (In the Guided Workflow, default settings are used)
- **Treatment of Control probes:** The control probes are included while performing normalization.
- **Empty Cells:** Not Applicable.
- **Sequence of events:** The sequence of events involved in the processing of the text data files is: Thresholding→log transformation→Normalization→Baseline Transformation

10.3 Guided Workflow steps

Summary report (Step 1 of 8): The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized

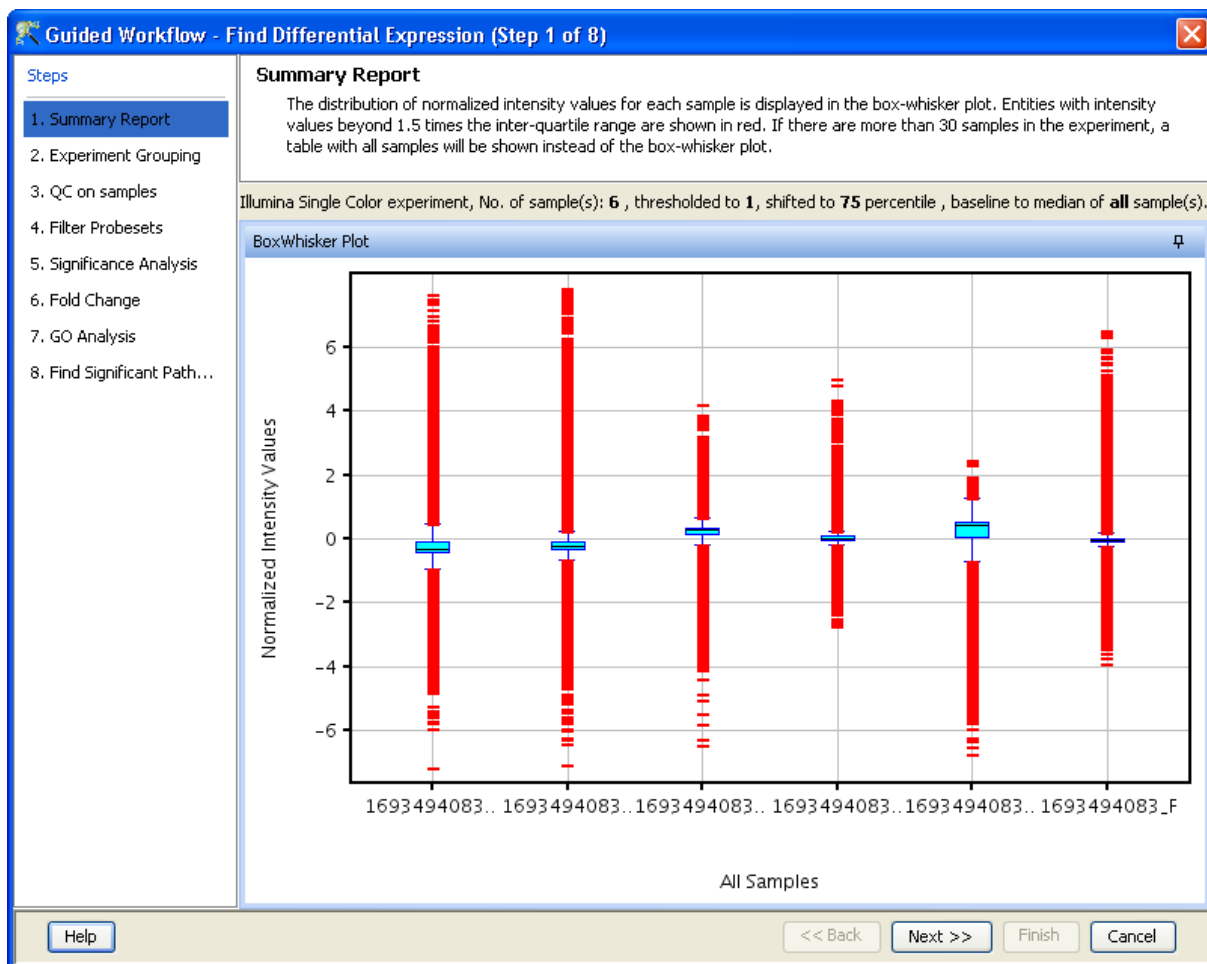


Figure 10.7: Summary Report



Expression values on the Y axis. An information message on the top of the wizard shows the number of samples in the file and the sample processing details. By default, the *Guided Workflow* does a thresholding of the signal values to 1. It then normalizes the data to 75th percentile and performs baseline transformation to median of all samples. If the number of samples are more than 30, they are only represented in a tabular column. On clicking the *Next* button it will proceed to the next step and on clicking *Finish*, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. On doing a right click, the options of invert selection is displayed and on clicking the same the selection is inverted i.e., all the probes except the selected ones are highlighted in green. Figure 10.7 shows the Summary report with box-whisker plot.

In the *Guided Workflow*, these default parameters cannot be changed. To choose different parameters use *Advanced Analysis*.

Experiment Grouping (Step 2 of 8): On clicking *Next*, the *Experiment Grouping* window appears which is the 2nd step in the **Guided Workflow**. It requires parameter values to be defined to

group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the **Add parameter** button. Parameter values can be assigned by first selecting the desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on **Clear**. Press **OK** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Guided Workflow**. The other parameters can be used in the **Advanced Analysis**.





Note: The *Guided Workflow* does not proceed further without grouping information.

Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file  icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

```
A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50
```

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file  icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click → *Properties* → *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click → *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

Note: The *Guided Workflow* by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the guided wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures 10.8 and 10.9 respectively.

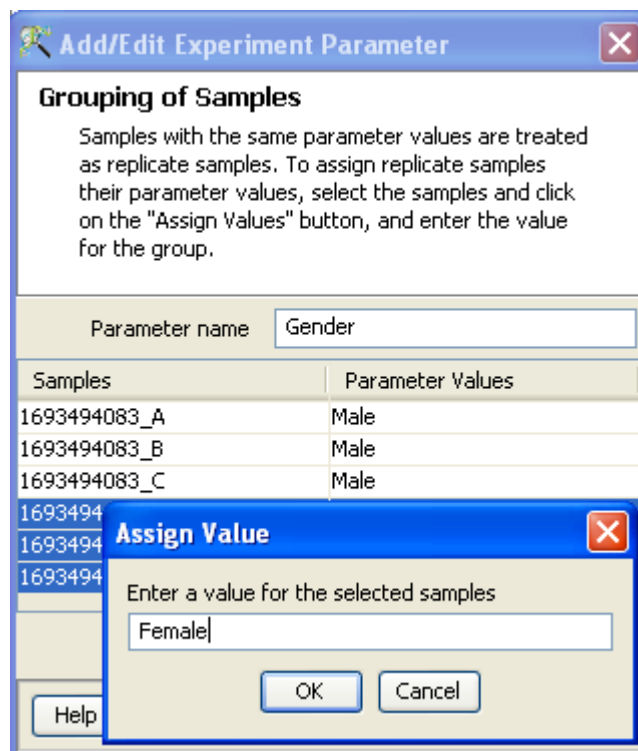


Figure 10.8: Experiment Grouping

Quality Control (Step 3 of 8): The 3rd step in the Guided workflow is the QC on samples which is displayed in the form of four tiled windows. They are as follows:

- Correlation coefficients table and Experiment grouping tabs
- Correlation coefficients plot
- PCA scores.
- Legend

QC on Samples generates four tiled windows as seen in Figure 10.10.

The views in these windows are lassoed i.e., selecting the sample in any of the view highlights the sample in all the views.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in two forms, one in textual form as a correlation table and other in visual form as a heatmap. The heatmap is colorable by *Experiment Factor* information via Right-Click→Properties. The intensity levels in the heatmap can also be customized here. The *Experiment Grouping* information is present along with the correlation table, as an additional tab.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X,

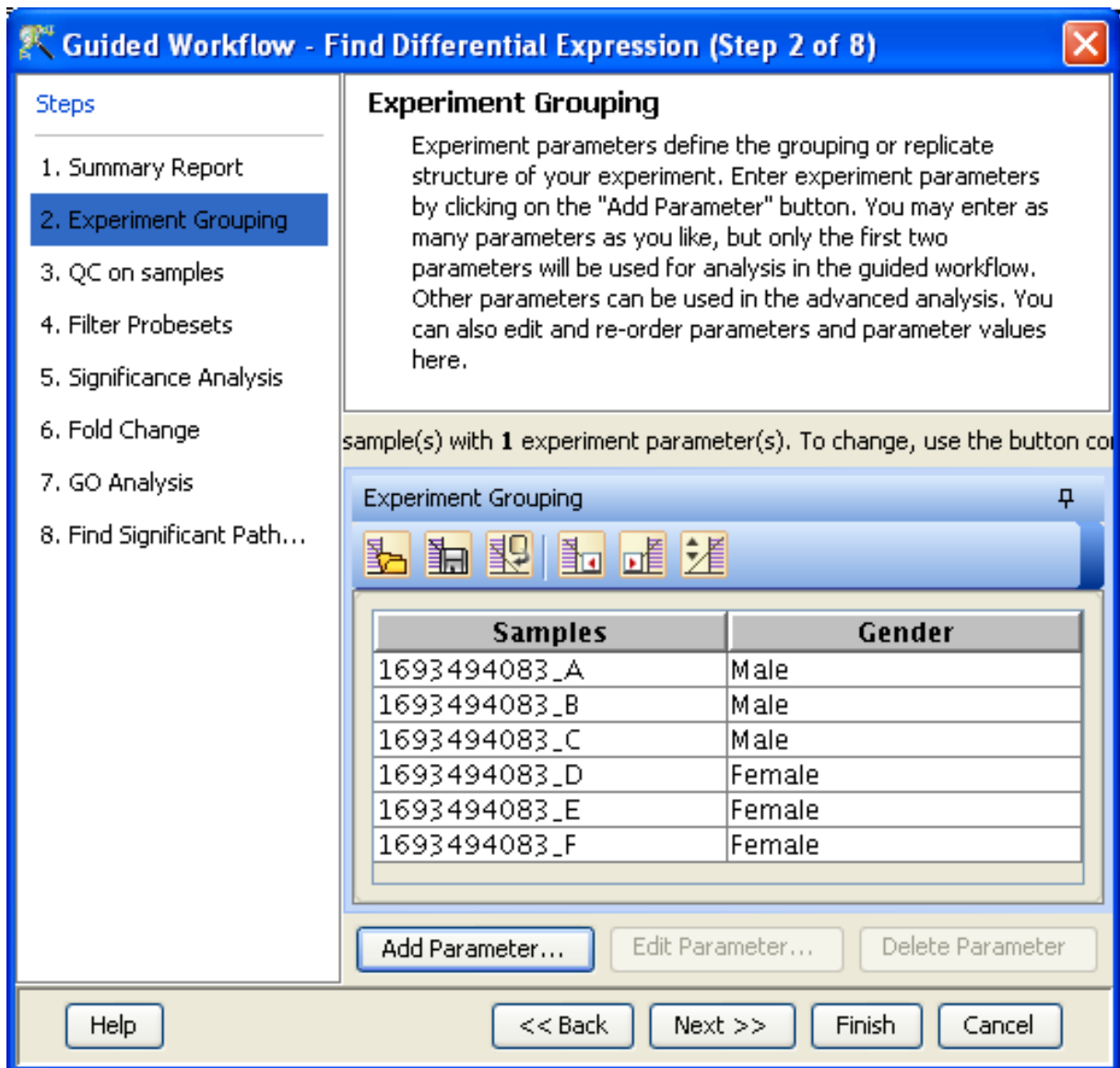


Figure 10.9: Edit or Delete of Parameters

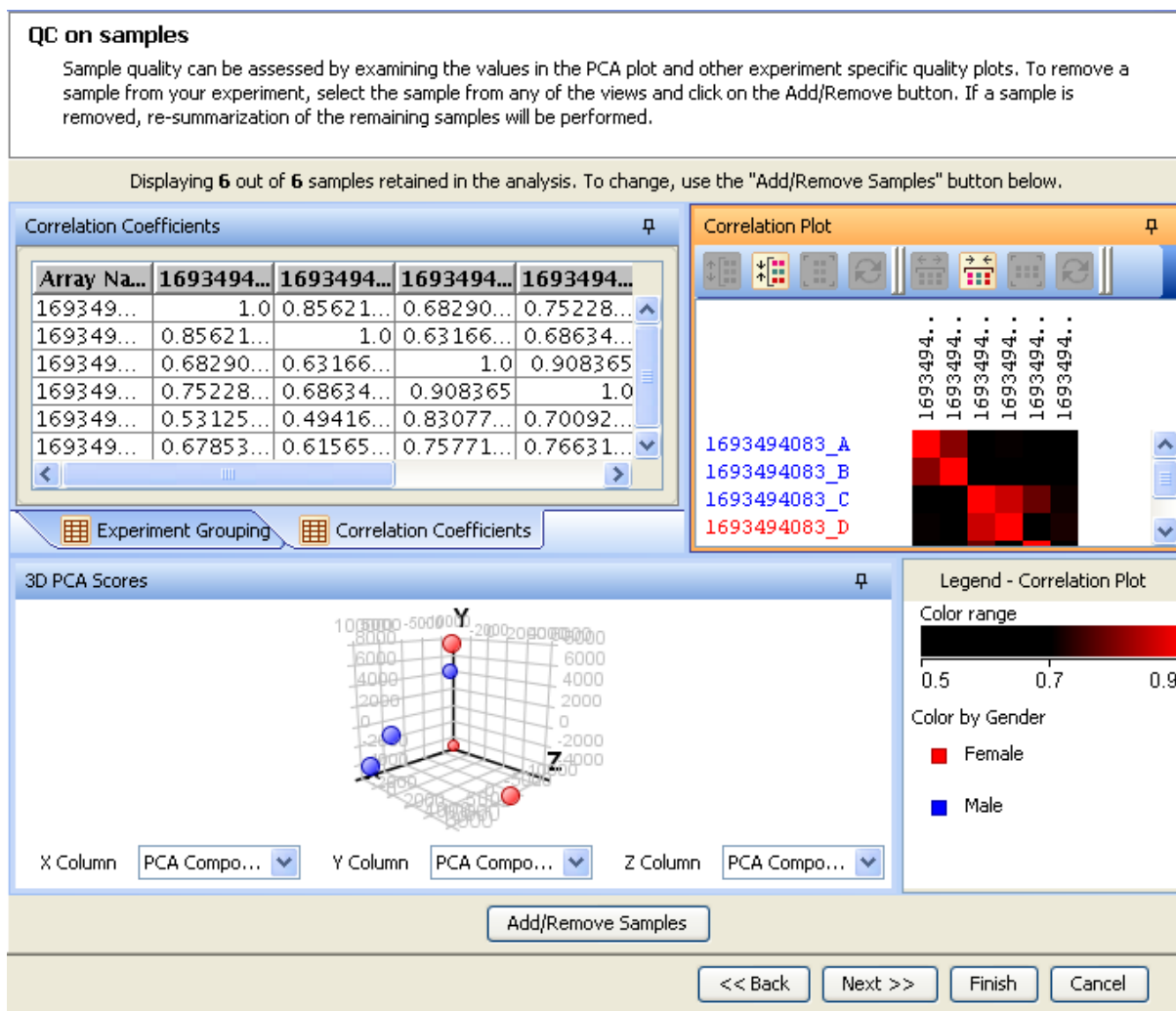


Figure 10.10: Quality Control on Samples

Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, normalization as well as baseline transformation is performed again on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

Filter Probesets (Step 4 of 8): In this step, the entities are filtered based on their flag values P(present), M(marginal) and A(absent). Only entities having the present and marginal flags in at least 1 sample are displayed as a profile plot. The selection can be changed using *Rerun Filter* option. The flag values are based on the Detection p-values columns present in the data file. Values below 0.06 are considered as Absent, between 0.06-0.08 are considered as Marginal and values above 0.08 are

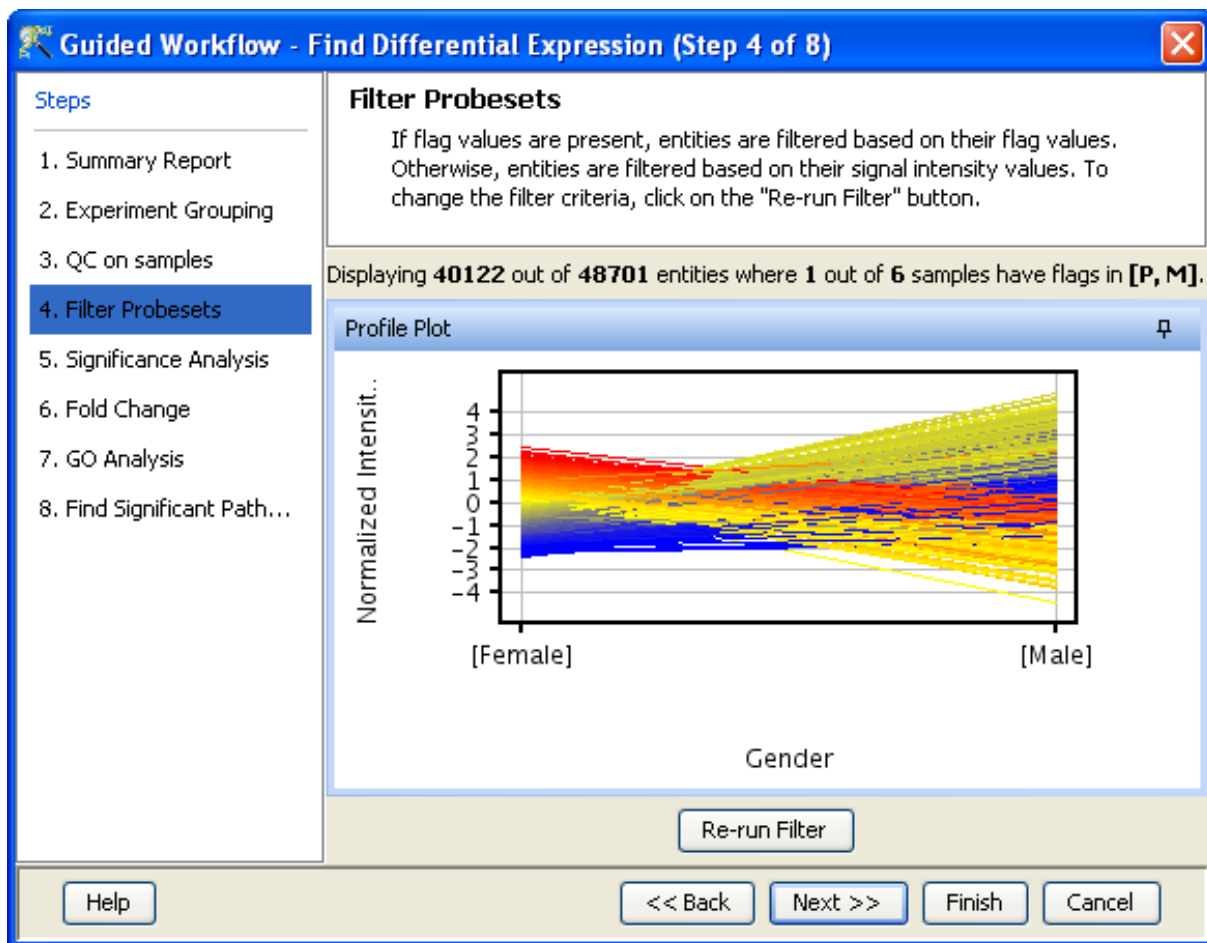


Figure 10.11: Filter Probesets-Single Parameter

considered as Present. To choose a different set of p-values representing Present, Marginal and Absent, go to the *Advanced Workflow*. The plot is generated using the normalized signal values and samples grouped by the active interpretation. Options to customize the plot can be accessed via the Right-click menu. An *Entity List*, corresponding to this filtered list, will be generated and saved in the Navigator window. The Navigator window can be viewed after exiting from *Guided Workflow*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Newer annotations can be added and existing ones removed using the *Configure Columns* button. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The cutoff for filtering can be changed using the *Rerun Filter* button. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. Double click on *Profile Plot* opens up an entity inspector giving the annotations corresponding to the selected profile. The information message on the top shows the number of entities satisfying the flag values.

Figures 10.11 and 10.12 are displaying the profile plot obtained in situations having a single and two parameters. Re-run option window is shown in 12.17

Significance analysis (Step 5 of 8): Depending upon the experimental grouping, **GeneSpring GX** performs either T-test or ANOVA. The tables below describe broadly the type of statistical test

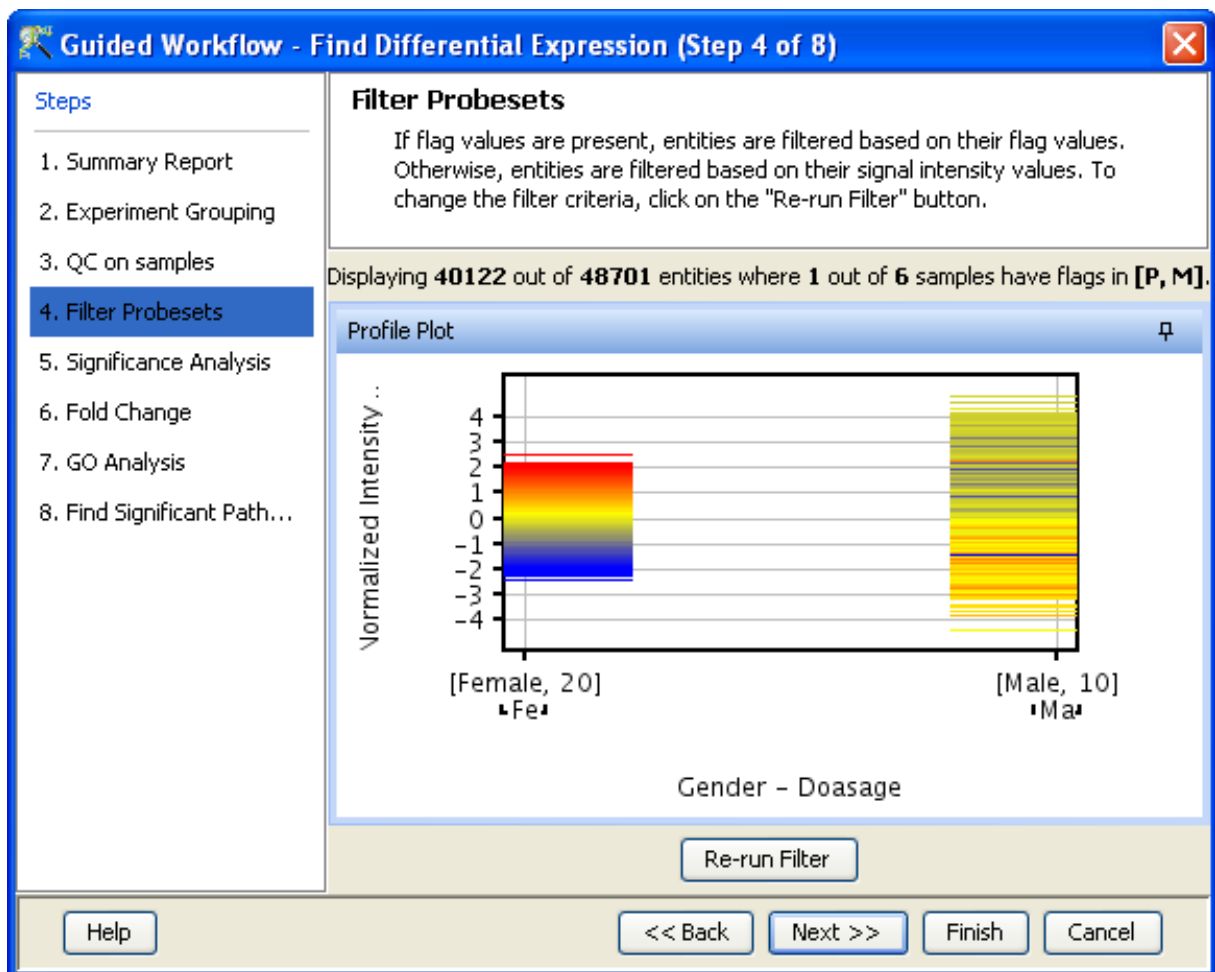


Figure 10.12: Filter Probesets-Two Parameters



Figure 10.13: Rerun Filter

performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I*, has 2 groups, the normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 10.1: Sample Grouping and Significance Tests I

- **Example Sample Grouping II:** In this example, only one group, the tumor, is present. T-test against zero will be performed here.

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 10.2: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When 3 groups are present (normal, tumor1 and tumor2) and one of the groups (tumor2 in this case) does not have replicates, statistical analysis cannot be performed. However if the condition tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 10.3: Sample Grouping and Significance Tests III

- **Example Sample Grouping IV:** When there are 3 groups within an interpretation, One-way ANOVA will be performed.
- **Example Sample Grouping V:** This table shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 10.4: Sample Grouping and Significance Tests IV

Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 10.5: Sample Grouping and Significance Tests V

- **Example Sample Grouping VI:** In this table, a two-way ANOVA will be performed.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 10.6: Sample Grouping and Significance Tests VI

- **Example Sample Grouping VII:** In the example below, a two-way ANOVA will be performed and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Statistical Tests: T-test and ANOVA

- **T-test: T-test unpaired** is chosen as a test of choice with a kind of experimental grouping shown in Table 1. Upon completion of T-test the results are displayed as three tiled windows.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 10.7: Sample Grouping and Significance Tests VII

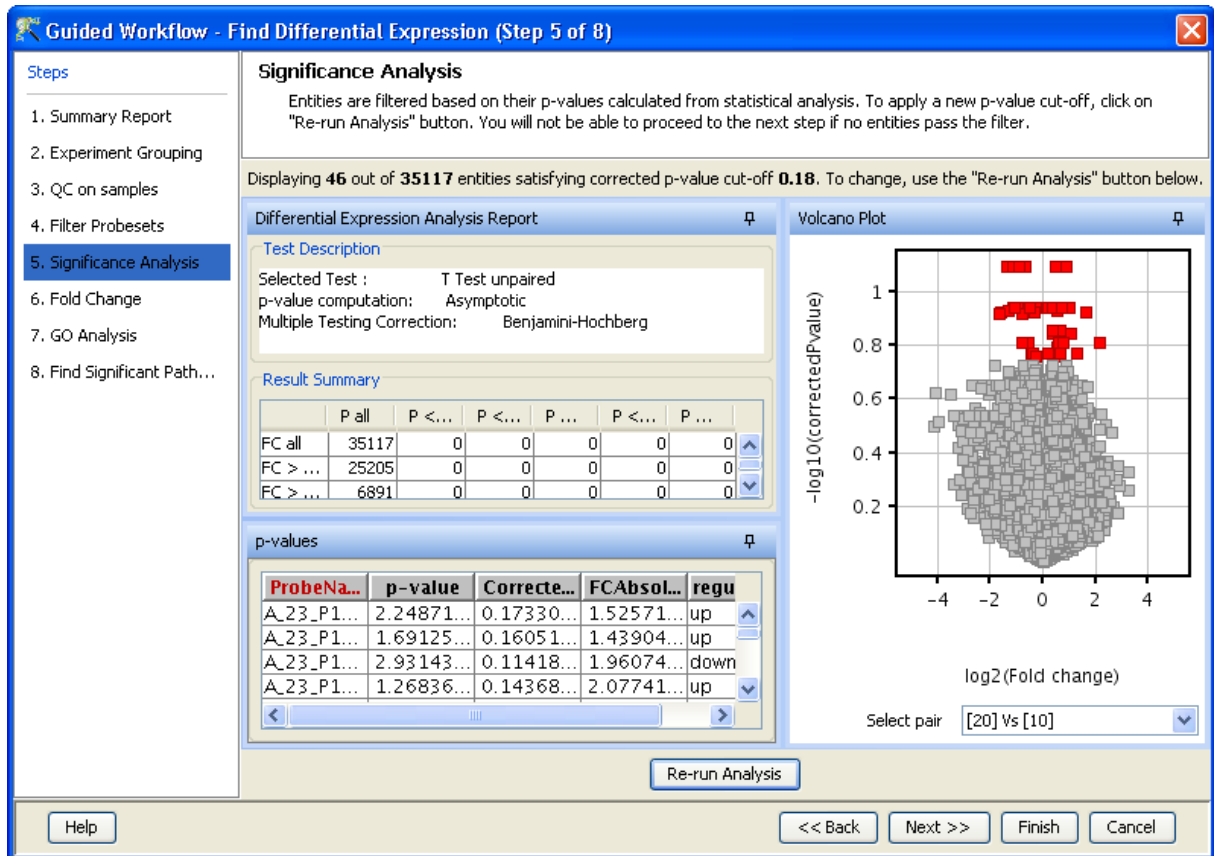


Figure 10.14: Significance Analysis-T Test

- A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *Regulation*.
- *Differential expression analysis report* mentioning the Test description i.e. test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic* or *Permutative*).

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

- **Analysis of variance (ANOVA):** ANOVA is chosen as a test of choice under the experimental

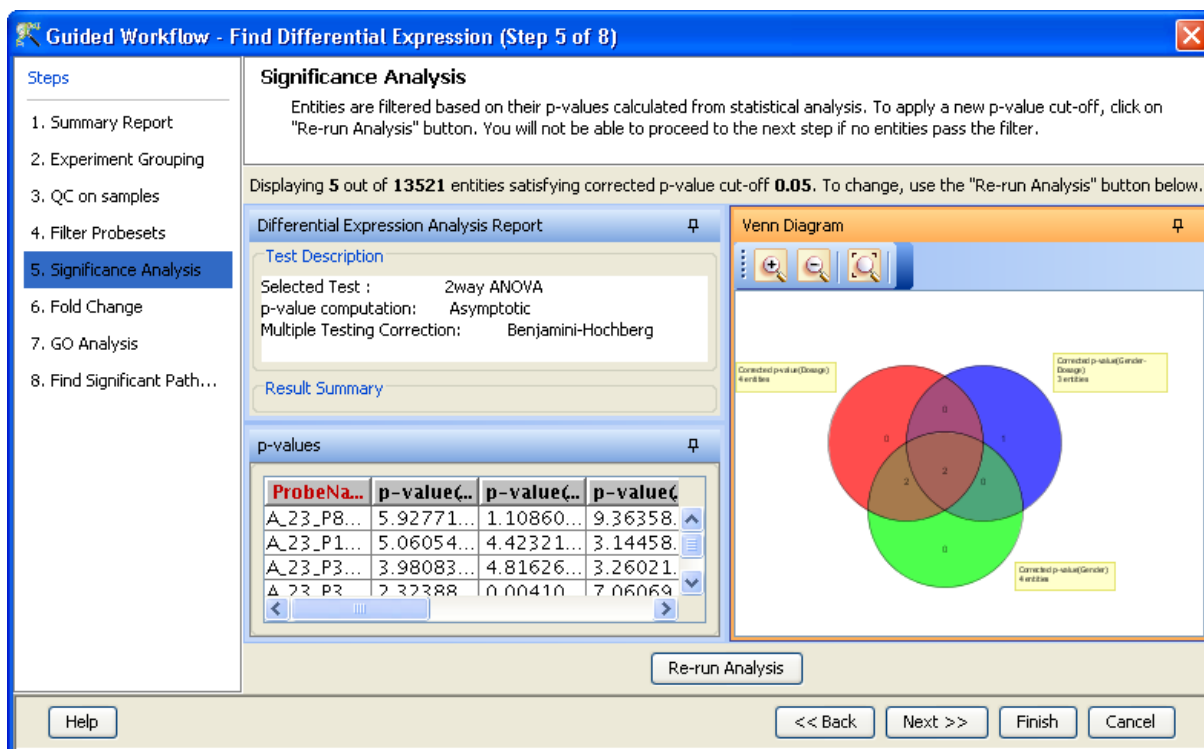


Figure 10.15: Significance Analysis-Anova

grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:

- A *p-value table* consisting of probe names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.
- *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and p-value computation type (*Asymptotic* or *Permutative*).
- *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

Special case: In situations when samples are not associated with at least one possible permutation of conditions (like Normal at 50 min and Tumor at 10 min mentioned above), no p-value can be computed and the **Guided Workflow** directly proceeds to **GO analysis**.

Fold-change (Step 6 of 8): **Fold change analysis** is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change

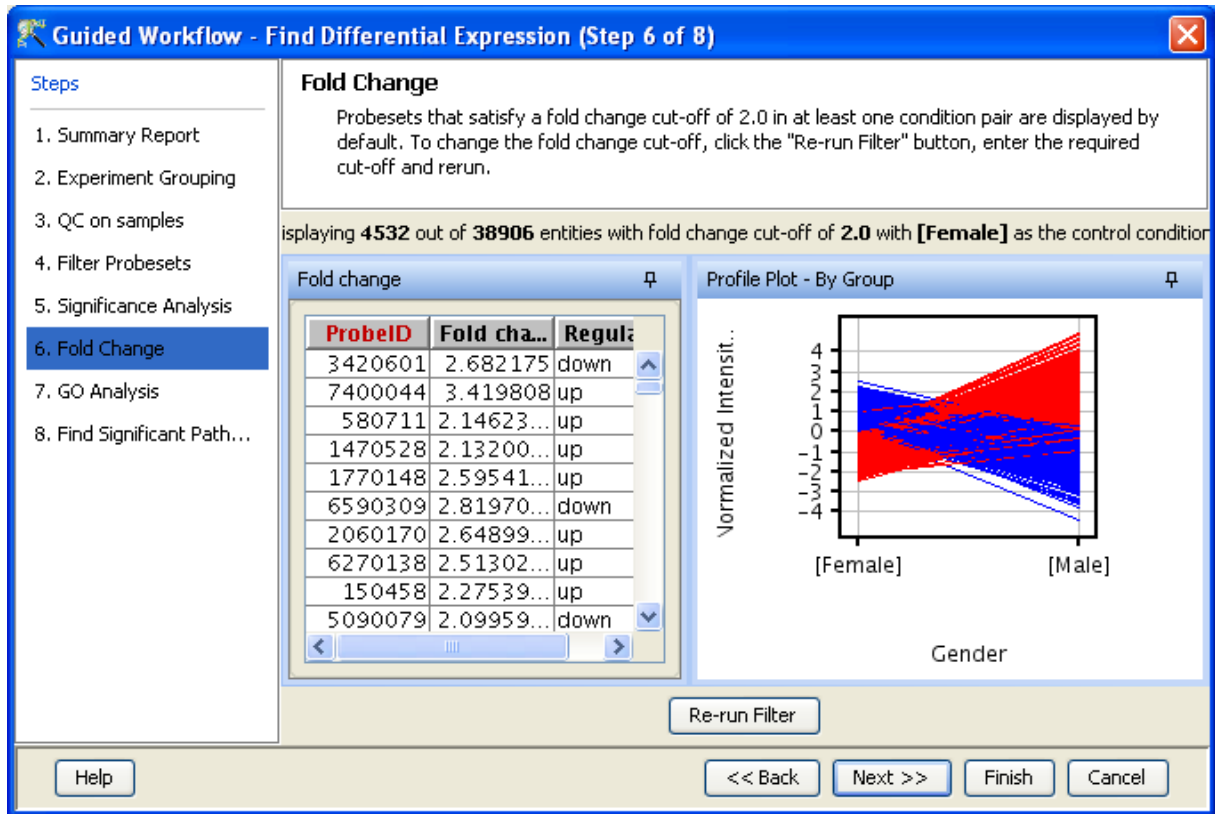


Figure 10.16: Fold Change

analysis. The wizard shows a table consisting of 3 columns: Probe Names, Fold change value and regulation (up or down). The regulation column depicts which one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using **Re-run Filter**. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1. A profile plot is also generated. Upregulated entities are shown in red. The color can be changed using the Right-click → *Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cutoff in the experiment Navigator.

Note: Fold Change step is skipped and the *Guided Workflow* proceeds to the *GO Analysis* in case of experiments having 2 parameters.

Fold Change view with the spreadsheet and the profile plot is shown in Figure 10.16.

Gene Ontology analysis (Step 7 of 8): The *GO Consortium* maintains a database of controlled vocabularies for the description of molecular function, biological process and cellular location of gene products. The GO terms are displayed in the Gene Ontology column with associated *Gene Ontology Accession* numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more

parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

The GO analysis wizard shows two tabs comprising of a spreadsheet and a *GO tree*. The *GO Spreadsheet* shows the *GO Accession* and *GO terms* of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However, the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.1 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button is also provided on the GO tree panel to search using some keywords

Note : In **GeneSpring GX** GO analysis implementation, all the three component: Molecular Function, Biological Processes and Cellular location are considered together.

On finishing the GO analysis, the *Advanced Workflow* view appears and further analysis can be carried out by the user. At any step in the Guided workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any) and the *Advanced Workflow* view appears.

Find Significant Pathways (Step 8 of 8): This step in the Guided Workflow finds relevant pathways from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the Fold Change (step 6 of 8). This view shows two tables-

- The Significant Pathways table shows the names of the pathways as well as the number of nodes and entities in the pathway and the p-values. It also shows the number of entities that are similar to the pathway and the entity list. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.
- The Non-significant Pathways table shows the pathways in the tool that do not have a single entity in common with the ones in the given entity list.

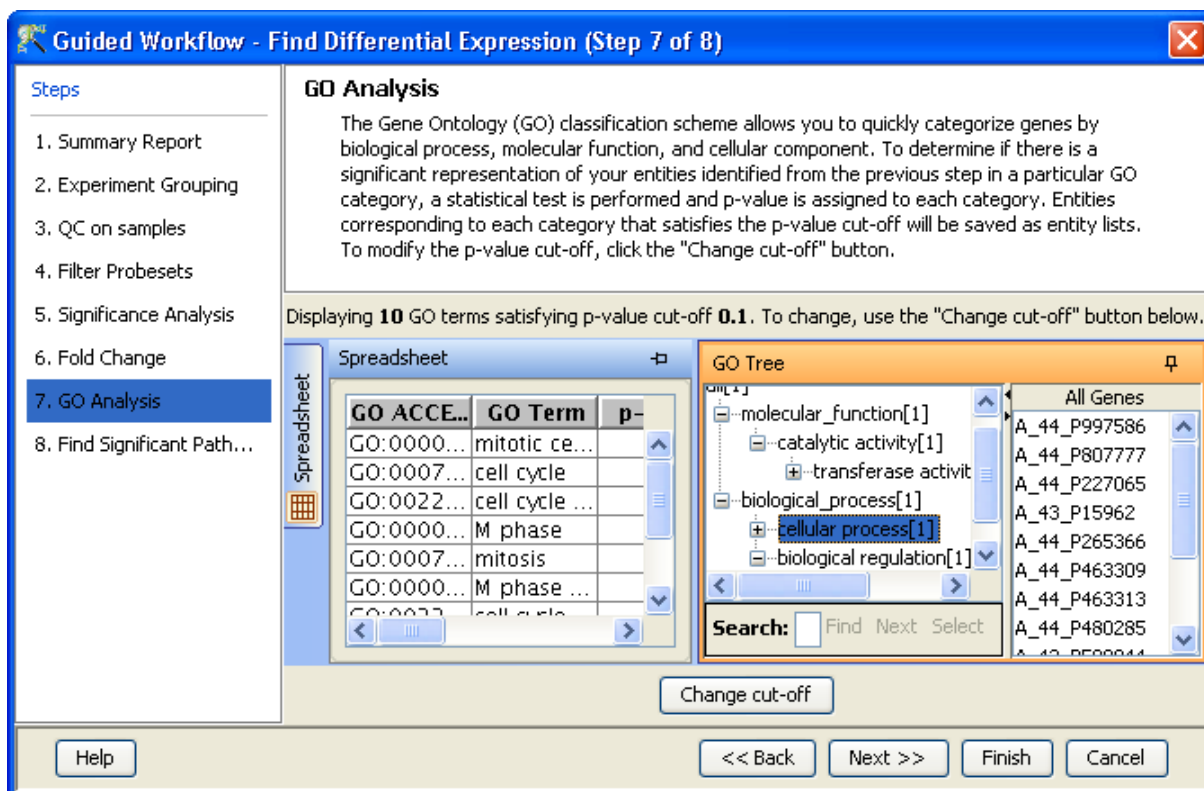


Figure 10.17: GO Analysis

The user has an option of changing the p-value cut-off (using *Change cutoff*) and also to save specific pathways using the *Custom Save* option. See figure 10.18. On clicking, *Finish* the main tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of the Guided Workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in the **Guided Workflow**, on clicking *Finish*, the analysis stops at that step (creating an entity list if any).

Note: In case the user is using **GeneSpring GX** for the first time, this option will give results using the demo pathways. The user can upload the pathways of his/her choice by using the option *Import BioPAX pathways* under **Tools** in the **Menu** bar in the main tool window. Later instead of reverting to the Guided Workflow the user can use the option *Find Significant Pathways* in **Results Interpretation** under the same Workflow.

The default parameters used in the Guided Workflow is summarized below.

10.4 Advanced Workflow:

The *Advanced Workflow* offers a variety of choices to the user for the analysis. The detection p-value range can be selected to decide on Present and Absent calls, raw signal thresholding can be altered and

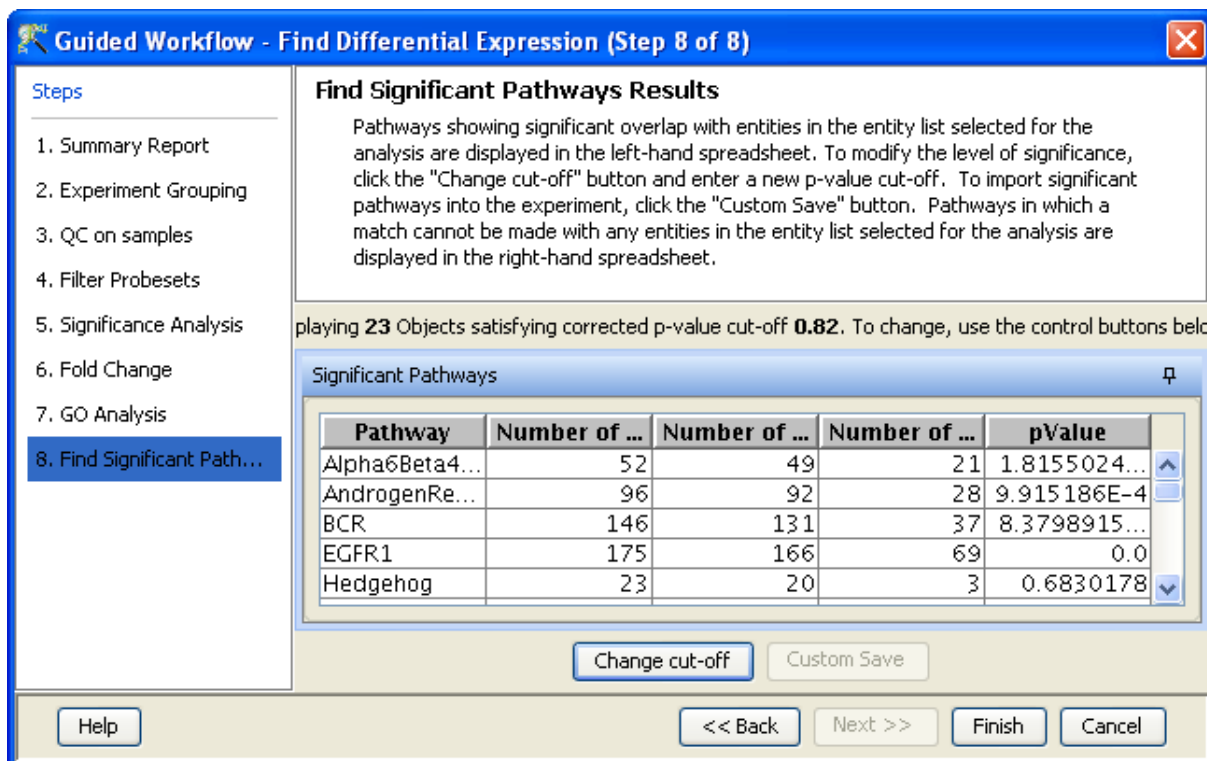


Figure 10.18: Fold Change

either Median Shift or Quantile Normalization can be chosen. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Advanced Workflow*, load the data as described earlier. In the *New Experiment Dialog*, choose the *Workflow Type* as Advanced. Click *OK* will open a new experiment wizard which then proceeds as follows:

1. **New Experiment (Step 1 of 5):** As in case of *Guided Workflow*, either data files can be imported or else pre-created samples can be used.
 - For loading new text files, use *Choose Files*.
 - If the txt files have been previously used in **GeneSpring GX** experiments *Choose Samples* can be used.

Step 1 of 3 of Experiment Creation, the 'Load Data' window, is shown in Figure 10.19.

2. **New Experiment (Step 2 of 5):** This step allows the user to determine the detection p-value range for Present and Absent flags. The Intermediate range will be taken as Marginal. The default values that are given for Present and Absent flags are 0.8 (lower cut-off) and 0.6 (upper cut-off) respectively. Step 2 of 3 of Experiment Creation, the Identify Calls Range window, is depicted in the Figure 10.20.
3. **New Experiment (Step 3 of 5):** Criteria for preprocessing of input data is set here. It allows the user to threshold raw signals to chosen values and to select normalization algorithms(Quantile, Percentile Shift, Scale and Normalize to control genes).

	Parameters	Parameter values
Expression Data Transformation	Thresholding	1.0
	Normalization	Shifted to 75th Percentile
	Baseline Transformation	Median of all samples
	Summarization	Not Applicable
Filter by		
1.Flags	Flags Retained	Present(P), Marginal(M)
2.Expression Values	(i) Upper Percentile cutoff	Not Applicable
	(ii) Lower Percentile cutoff	
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1
Find Significant Pathways	p-value cutoff	0.05

Table 10.8: Table of Default parameters for Guided Workflow

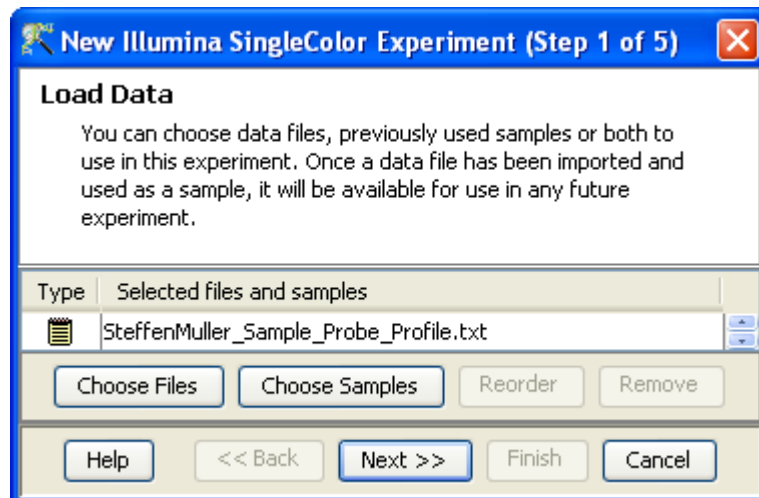


Figure 10.19: Load Data

- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

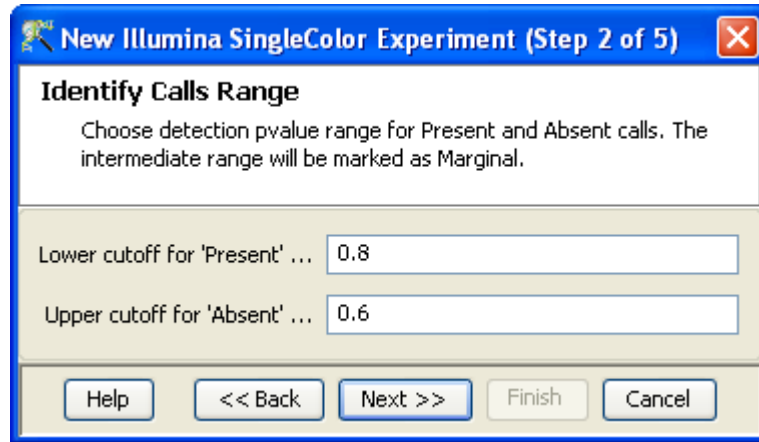


Figure 10.20: Identify Calls Range

- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the '*Assign Value*' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.

For details on the above normalization methods, refer to section [Normalization Algorithms](#).

Figure 10.21 shows the Step 3 of 5 of Experiment Creation.

Experiment (Step 4 of 5): If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the ***Choose Entities*** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click ***Finish***. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selection is over. See figure 10.22.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Experiment (Step 5 of 5): This step allows the user to perform baseline transformation. See figure 10.23. The baseline options include:

- ***Do not perform baseline***

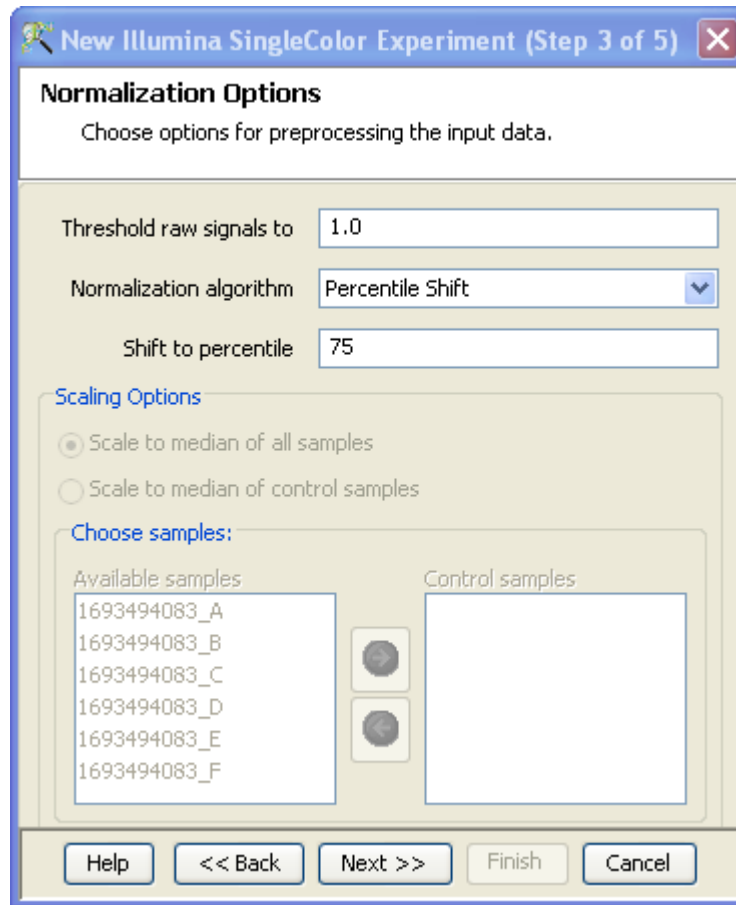


Figure 10.21: Preprocess Options

- **Baseline to median of all samples:** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- **Baseline to median of control samples:** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on **Assign value**. This opens up the **Choose Control Samples** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on **Ok**. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Once an experiment is created, the *Advanced Workflow* steps appear on the right hand side. Following is an explanation of the various workflow links:

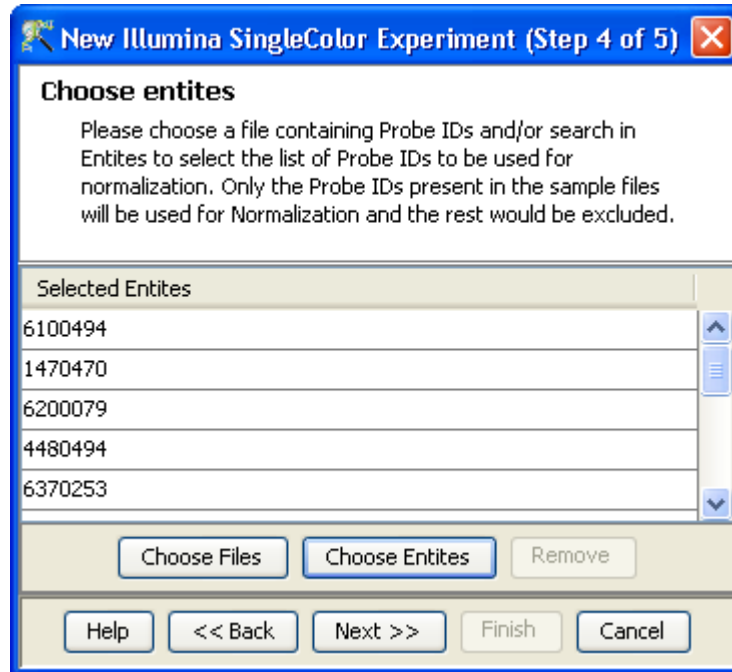


Figure 10.22: Choose Entities

10.4.1 Experiment Setup

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the Advanced Workflow, the method of analysis, the details of the algorithms used and the interpretation of results
- **Experiment Grouping:** Experiment parameters defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.

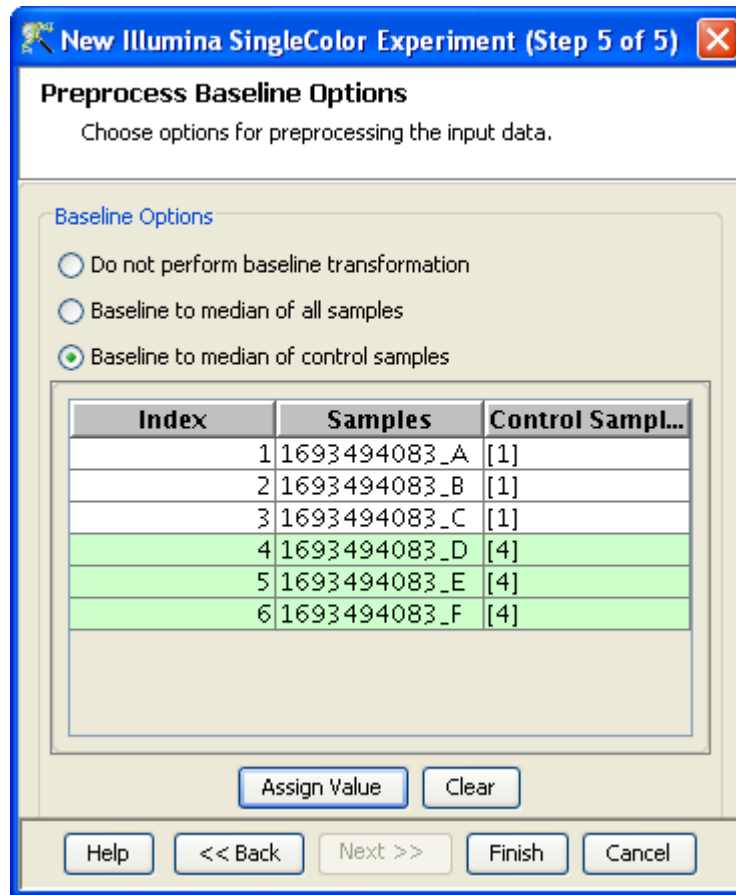


Figure 10.23: Preprocess Baseline Options

- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

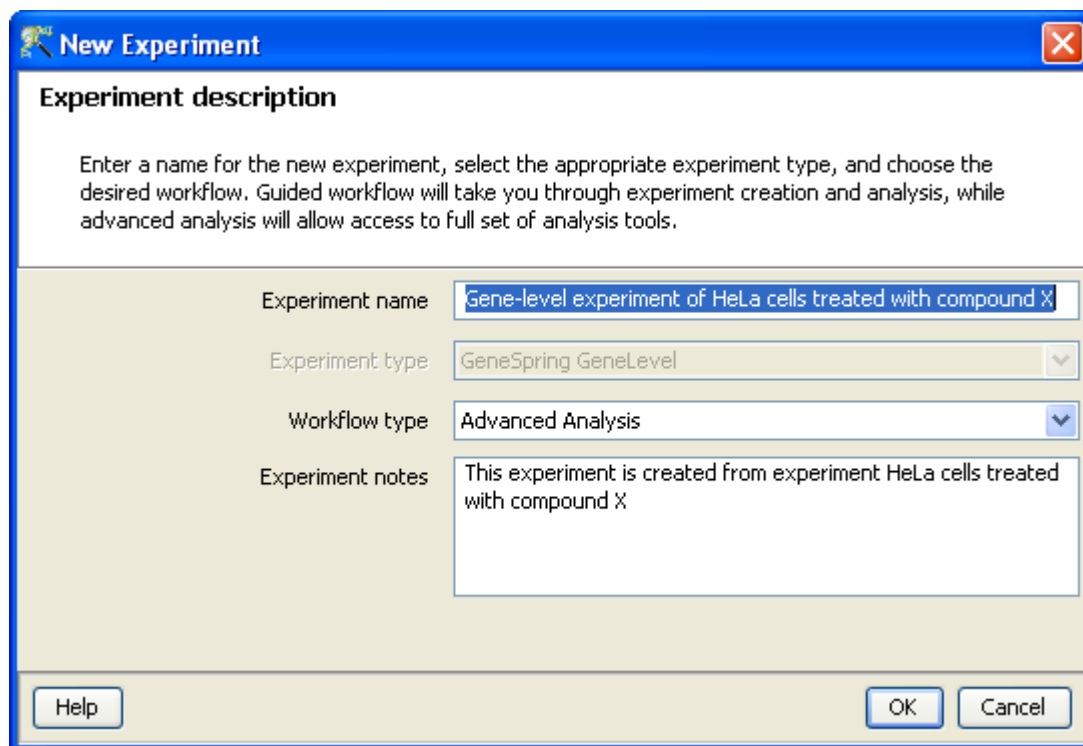


Figure 10.24: Gene Level Experiment Creation

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

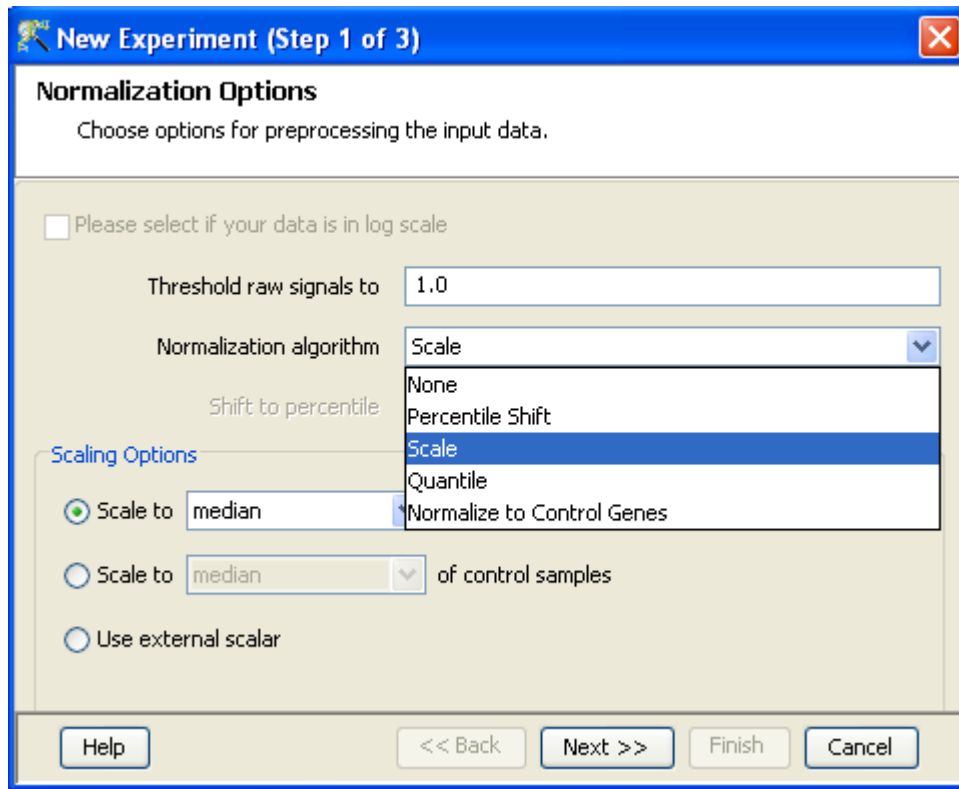


Figure 10.25: Gene Level Experiment Creation - Normalization Options

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

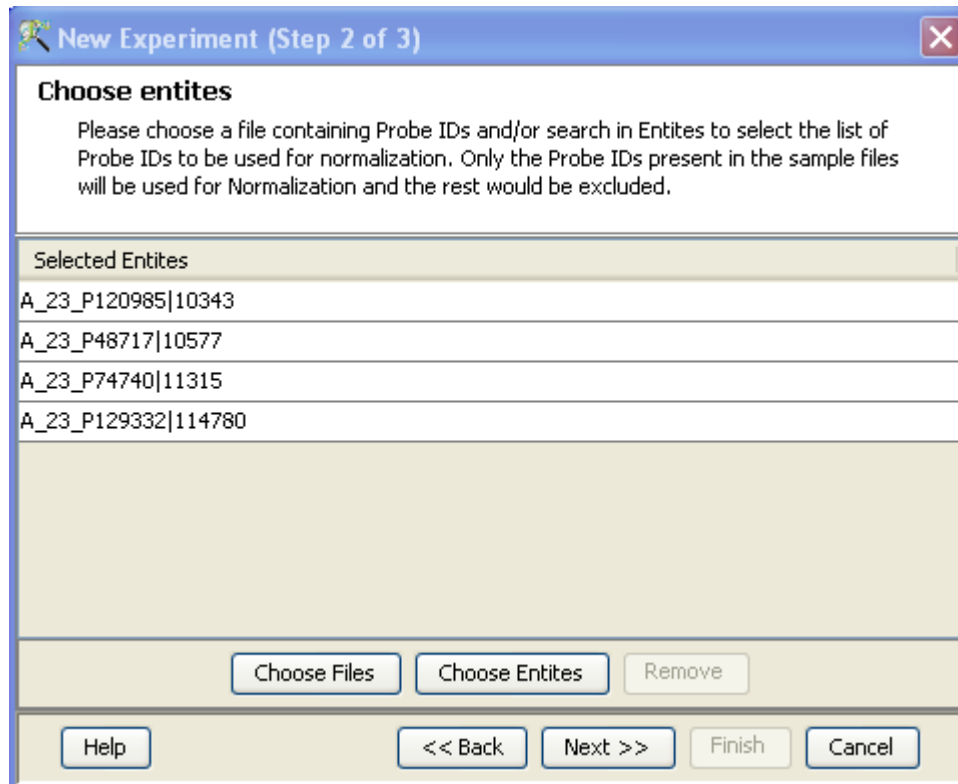


Figure 10.26: Gene Level Experiment Creation - Choose Entities

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

10.4.2 Quality control

- **Quality Control on samples:**

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows four tiled windows:

- Correlation plots and Correlation coefficients
- Experiment grouping
- PCA scores
- Legend

Figure 10.28 has the 4 tiled windows which reflect the QC on samples.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

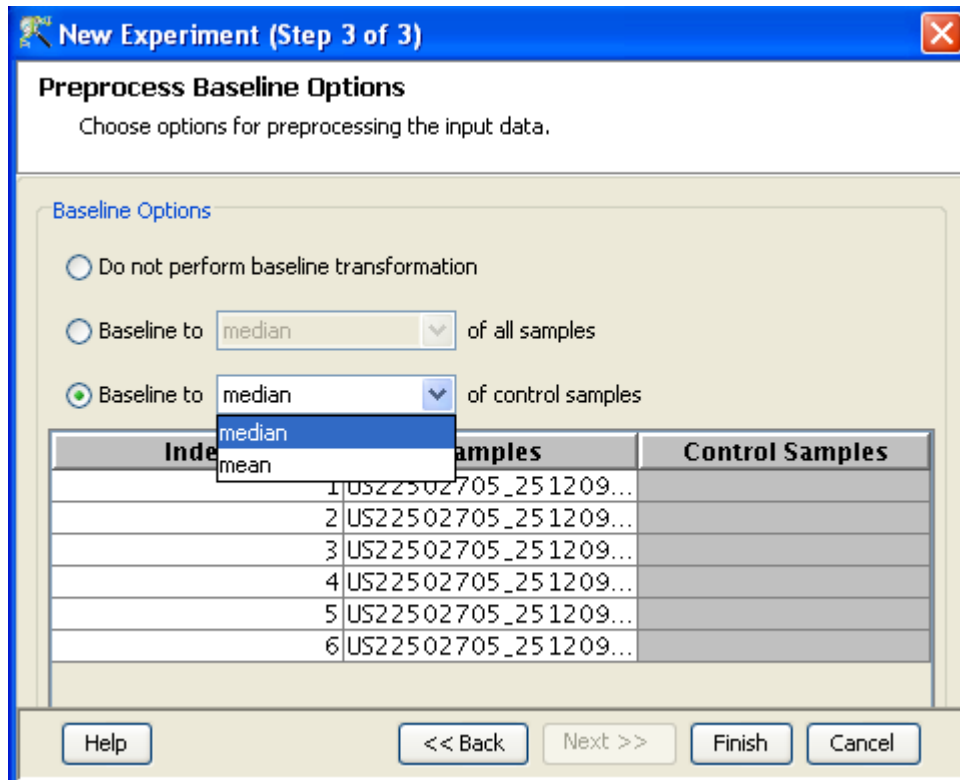


Figure 10.27: Gene Level Experiment Creation - Preprocess Baseline Options

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$

The heatmap is colorable by Experiment Factor information via Right-Click → Properties. Similarly, the intensity levels in the heatmap are also customizable.

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

Experiment Grouping shows the parameters and parameter values for each sample.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click** → **Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

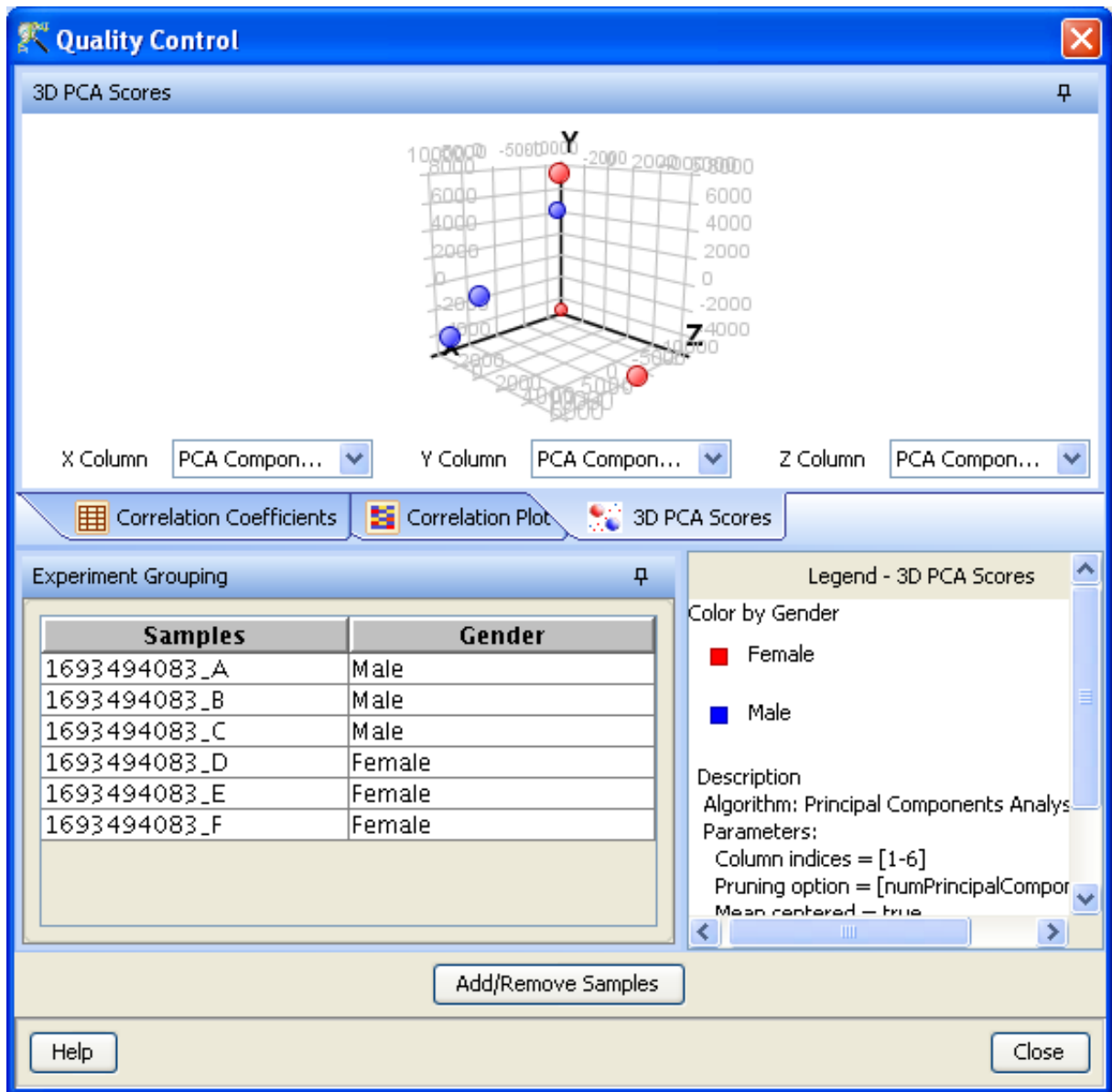


Figure 10.28: Quality Control

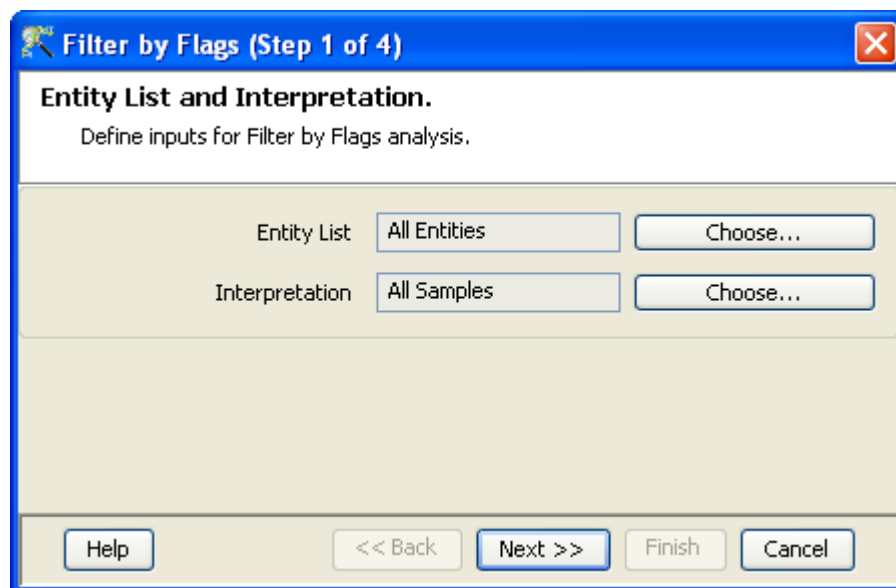


Figure 10.29: Entity list and Interpretation

The fourth window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on *OK* to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags:** In this step, the entities are filtered based on their flag values, the P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new experiment (Step 2 of 3) are taken into consideration while filtering the entities. The filtration is done in 4 steps:
 1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.
 2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.
 3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window (See Figure 10.31).
 4. Step 4 of 4: Click *Next* to annotate and save the entity list (See Figure 10.32).
- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on [Filter Probesets on Data Files](#)

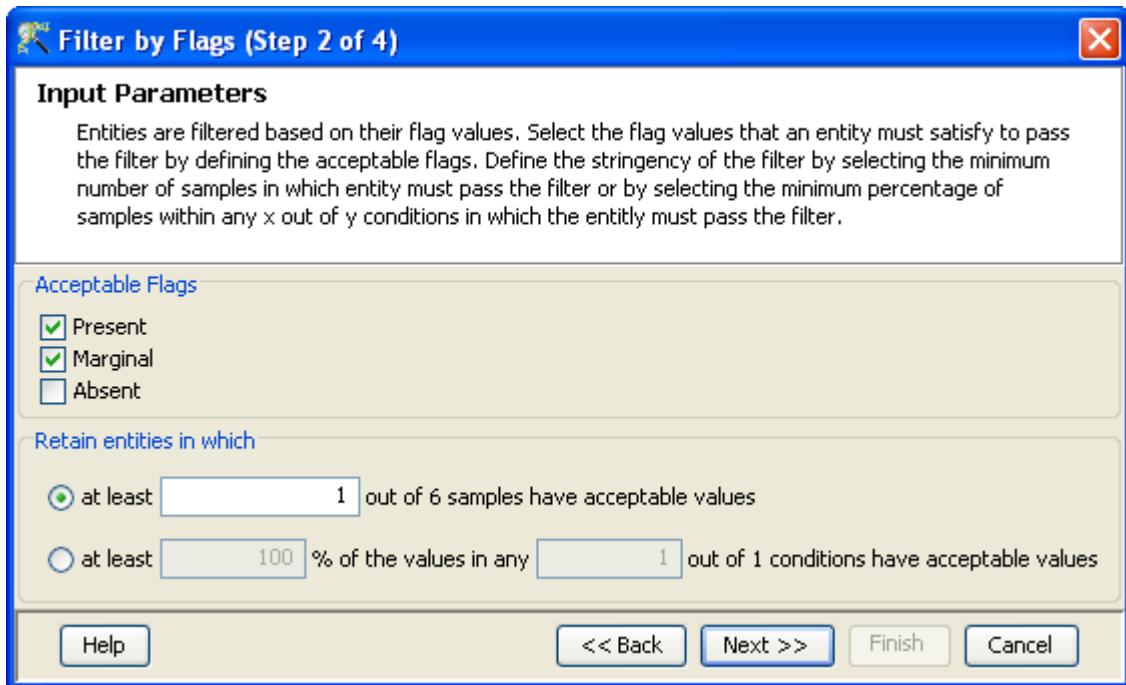


Figure 10.30: Input Parameters

- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

10.4.3 Analysis

- **Statistical Analysis**
For details refer to section [Statistical Analysis](#) in the advanced workflow.
- **Filter on Volcano Plot**
For details refer to section [Filter on Volcano Plot](#)
- **Fold Change**
For details refer to section [Fold Change](#)
- **Clustering**
For details refer to section [Clustering](#)
- **Find Similar Entities**
For details refer to section [Find Similar Entities](#)
- **Filter on Parameters**
For details refer to section [Filter on Parameters](#)

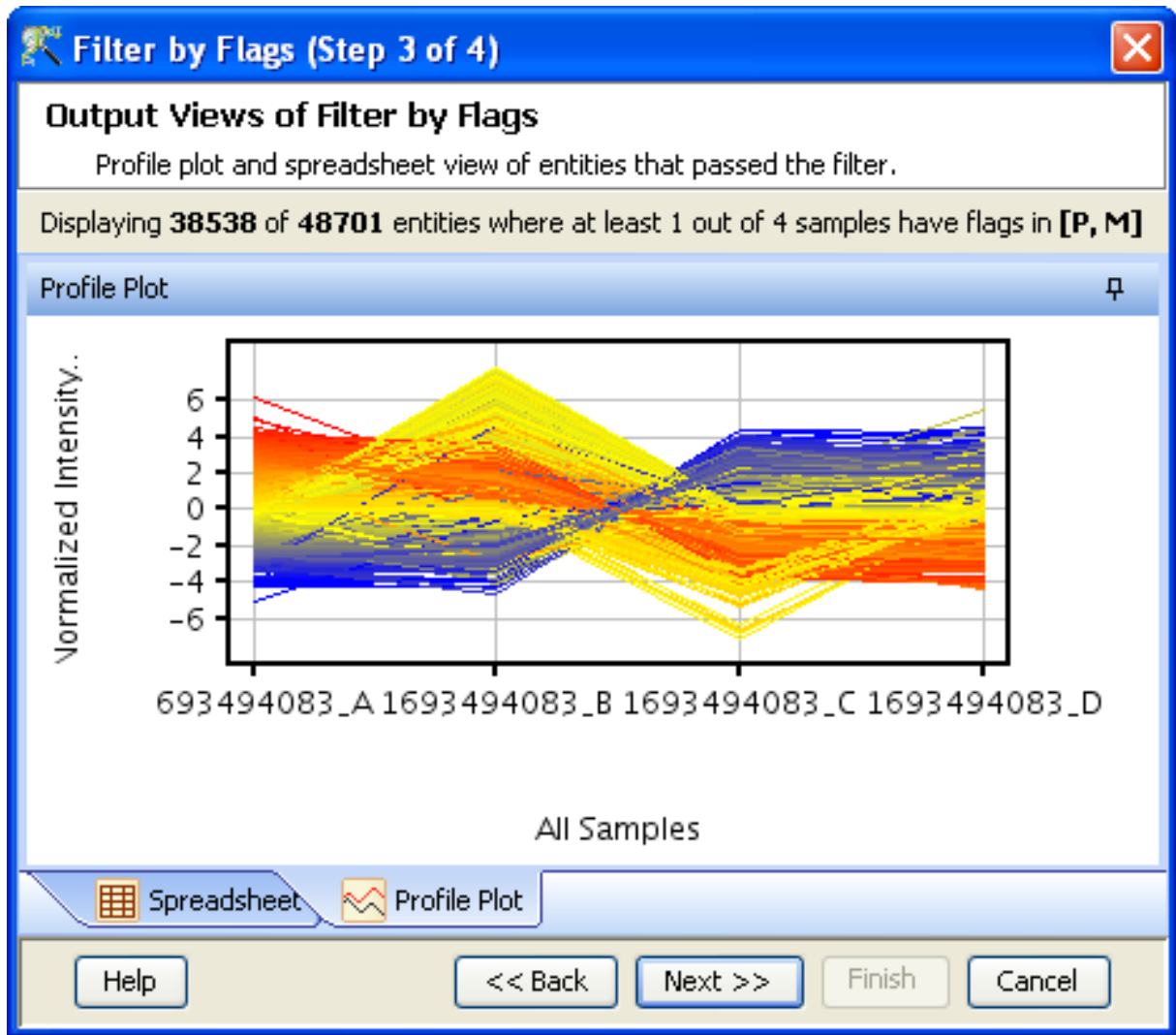


Figure 10.31: Output Views of Filter by Flags

- **Principal Component Analysis**

For details refer to section [PCA](#)

10.4.4 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

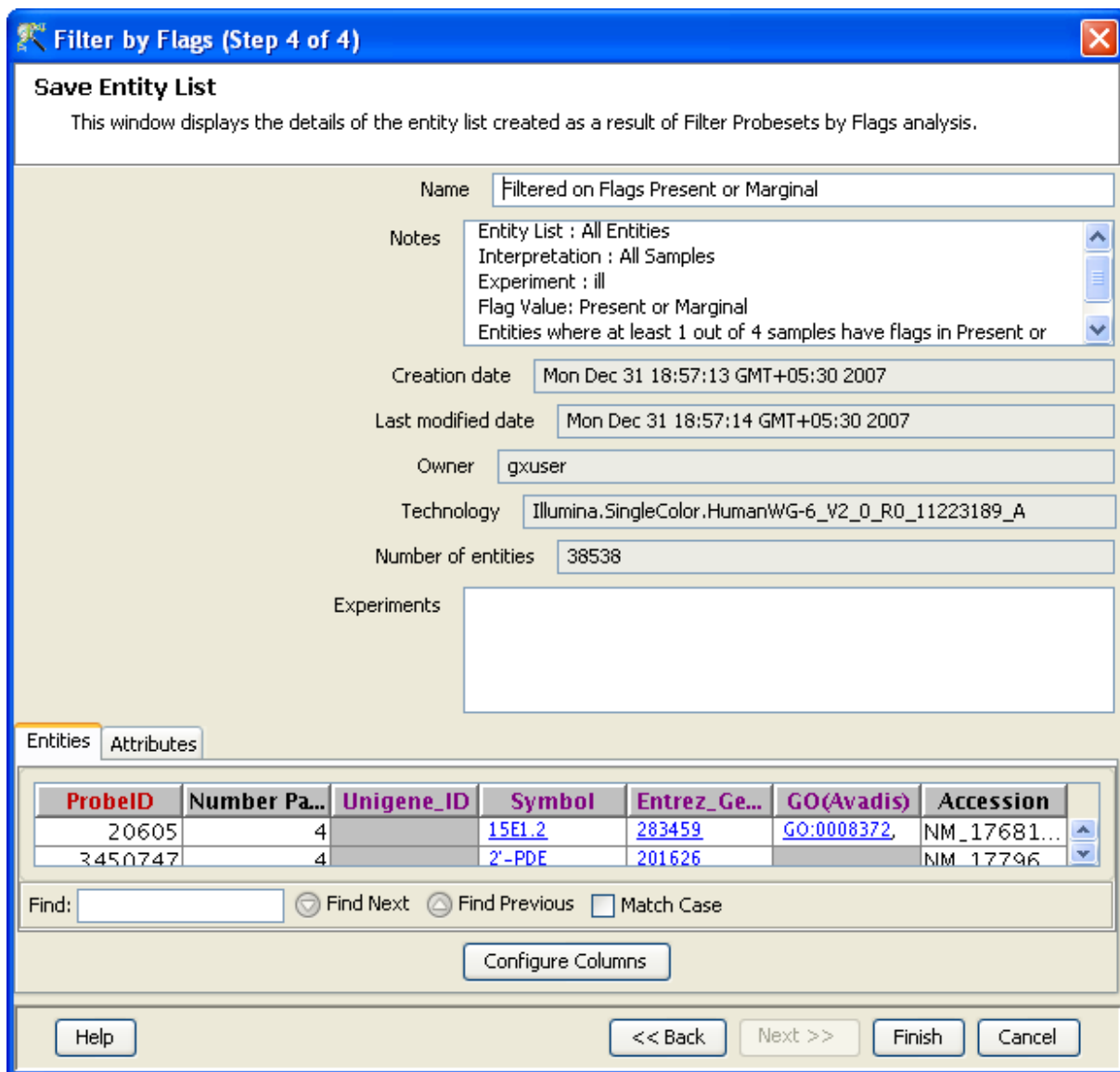


Figure 10.32: Save Entity List

10.4.5 Results

- **Gene Ontology (GO) analysis**

GO is discussed in a separate chapter called [Gene Ontology Analysis](#).

- **Gene Set Enrichment Analysis (GSEA)**

Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).

- **Gene Set Analysis (GSA)**

Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).

- **Pathway Analysis**

Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).

- **Find Similar Entity Lists**

This feature is discussed in a separate section called [Find Similar Entity Lists](#)

- **Find Significant Pathways**

This feature is discussed in a separate section called [Find Significant Pathways](#).

- **Launch IPA**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Import IPA Entity List**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Extract Interactions via NLP**

This feature is discussed in detail in the chapter [Pathway Analysis](#).

10.4.6 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)

- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)

- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)

- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

10.4.7 Illumina Custom Technology creation

The number of standard technologies available for Illumina in **GeneSpring GX** can be obtained from *Annotations* → *Create Technology* → *From Agilent Server*. Illumina projects can also be analyzed by creating a custom technology (*Annotations* → *Create Technology* → *Custom from file*) and then using the Generic Single Color workflow. This is done in either of the 2 cases:

- If you have projects created using Illumina technologies, which are not supported by the Illumina Single Color Importer.
- If you need additional annotation columns (over and above which comes when you do a GeneSpring Format export) from Genome Studio like the probe sequence, probe coordinates etc..

To create a Custom Technology using an Illumina Genome Studio project, follow the steps outlined below:

- Create a project in Bead Studio using either the .xml or the .bgx Content Descriptor file.
- Once a project is created, four spreadsheets-the Sample Probe profile, Sample Gene profile, Group Probe profile, and Group Gene profile are generated. These contain the Intensity values and some annotations. More annotations can be brought into these by going to Column Chooser in Genome Studio. This allows you to either show or hide additional columns. This file, either with or without additional annotations can be exported out as a text file.
- The file can be exported either in GeneSpring format by going to *File* → *Export in GeneSpring Format* or in a tab delimited text format by clicking on the Export Displayed Data to File icon in Genome Studio.
- These text files can then be imported into **GeneSpring GX** to create a Generic Single Color experiment. For details on creating a Generic Single Color experiment, refer to the Chapter 15 on [Creating Technology](#). In the process of technology creation, use the data file containing the annotations as both the data and annotation file.
- Proceed with the rest of the steps as usual.

Chapter 11

Analyzing Agilent Single Color Expression Data

GeneSpring GX supports Agilent Single Color technology. The data files are in .txt format and are obtained from Agilent Feature Extraction(FE) 8.5 and 9.5.3. When the data file is imported into **GeneSpring GX** the following columns get imported for the purpose of experiment creation: ControlType, ProbeName, Signal and Feature Columns.

An Agilent Single Color Workflow can be used if either a single color experiment is performed or if a two color experiment is performed but subsequent analysis requires the splitting of the channel into 2 individual channels. These 2 channels can then be treated as 2 single color samples. For the latter situation, see the section on [Analyzing Agilent Two Color data in Agilent Single Color Experiment Type](#)

The Agilent Single Color Workflow supports most of the Standard Agilent technologies. The Agilent custom arrays and the files from FE other than 8.5 and 9.5.3 can be analyzed by creating a Generic Single Color technology using the corresponding workflow. In order to do so, certain column markings should be indicated (which are automatically done with standard technologies). These details can be found in the section on [Custom Agilent Arrays](#), while the Generic Single Color technology creation is available in Chapter 15 in the section [Creating Technology](#)

11.1 Running the Agilent Single Color Workflow

Upon launching **GeneSpring GX** , the startup is displayed with 3 options.

- Create new project
- Open existing project

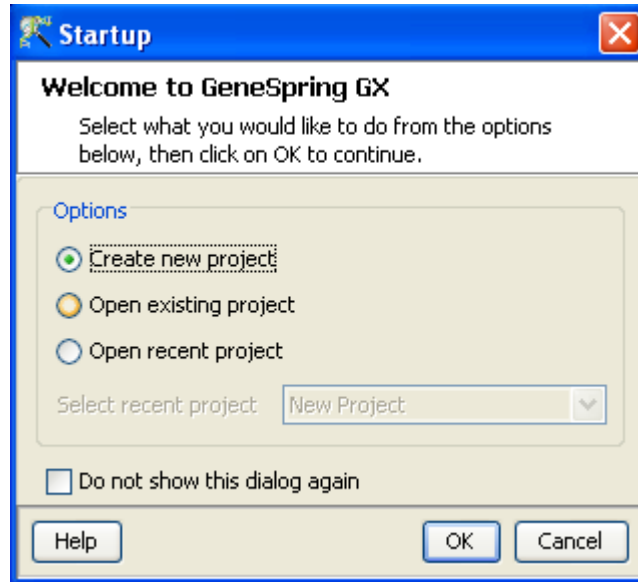


Figure 11.1: Welcome Screen

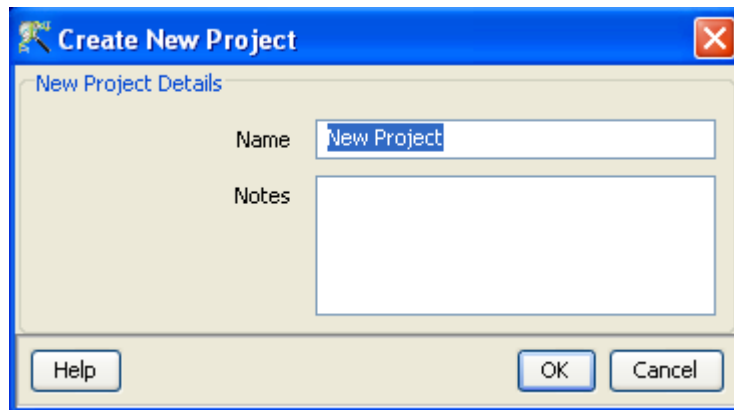


Figure 11.2: Create New project

- **Open recent project**

Either a new project can be created or a previously generated project can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on **OK** to proceed.

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment:** This allows the user to create a new experiment. (steps described below).

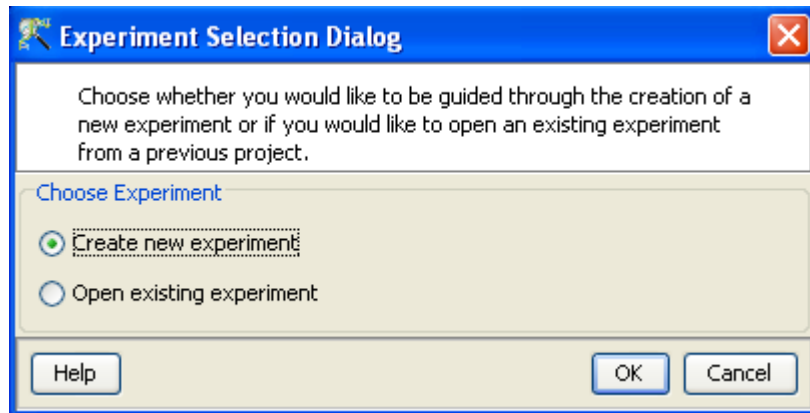


Figure 11.3: Experiment Selection

2. **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the experiment type gives the user the option to choose between the multiple experiment types namely Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Agilent miRNA, Generic Single Color, Generic Two Color, Pathway and RealTime-PCR experiment.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Guided Workflow**
2. **Advanced Analysis Workflow**

Guided Workflow is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Advanced Analysis**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples
3. Reorder

4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. On the other hand, a sample is created within **GeneSpring GX**, when it associates the data files with its appropriate technology (See the section on [Technology](#)). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the *Choose Samples* option. For selecting data files and creating an experiment, click on the *Choose File(s)* button, navigate to the appropriate folder and select the files of interest. Click on *OK* to proceed.

The technology specific for any chip type needs to be created or downloaded only once. Thus, upon creating an experiment of a specific chip type for the first time, **GeneSpring GX** prompts the user to download the technology from the update server. If the technology is not present, then **GeneSpring GX** creates it on the fly using user provided data identifiers. Annotations from a file can be added at any time by going to *Annotations*→*Update Technology Annotations*. If an experiment has been created previously with the same technology, **GeneSpring GX** then directly proceeds with experiment creation. Clicking on the *Choose Samples* button, opens a sample search wizard, with the following search conditions:

1. **Search field:** Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition:** Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on *Add* button, or can be removed from the list using *Remove* button.

Files can either be removed or reordered during the data loading step using the *Remove* or *Reorder* button.

Figures [11.4](#), [11.5](#), [11.6](#), [11.7](#) show the process of choosing experiment type, loading data, choosing samples and re-ordering the data files.

11.1.1 Analyzing Agilent Two Color data in Agilent Single Color Experiment Type

Essentially a Two Color technology can be used to analyze two samples within one slide or multiple samples in different arrays of a slide. This can be done in the following experimental designs: Imagine you have the

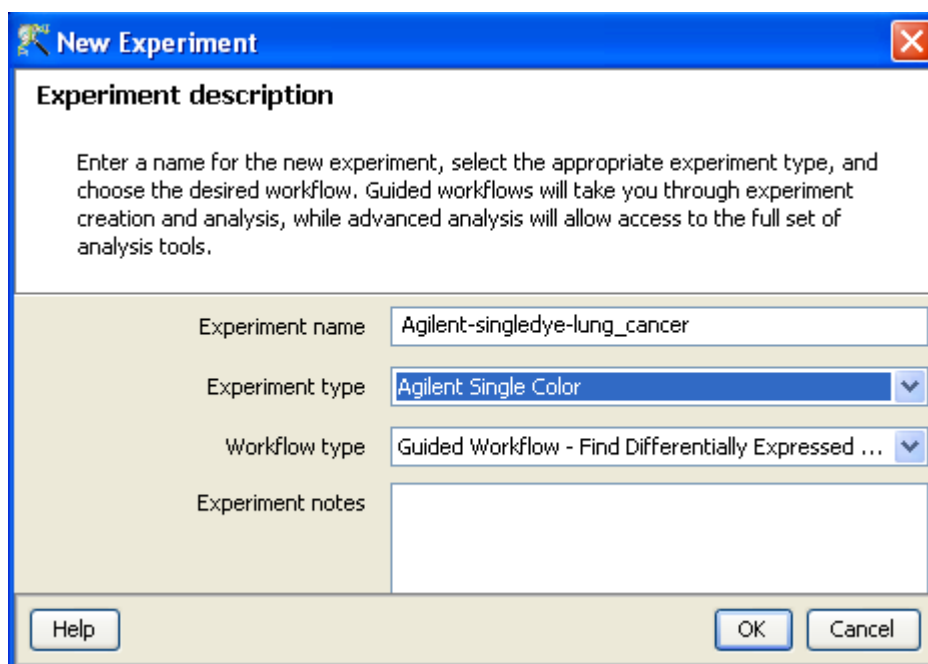


Figure 11.4: Experiment Description

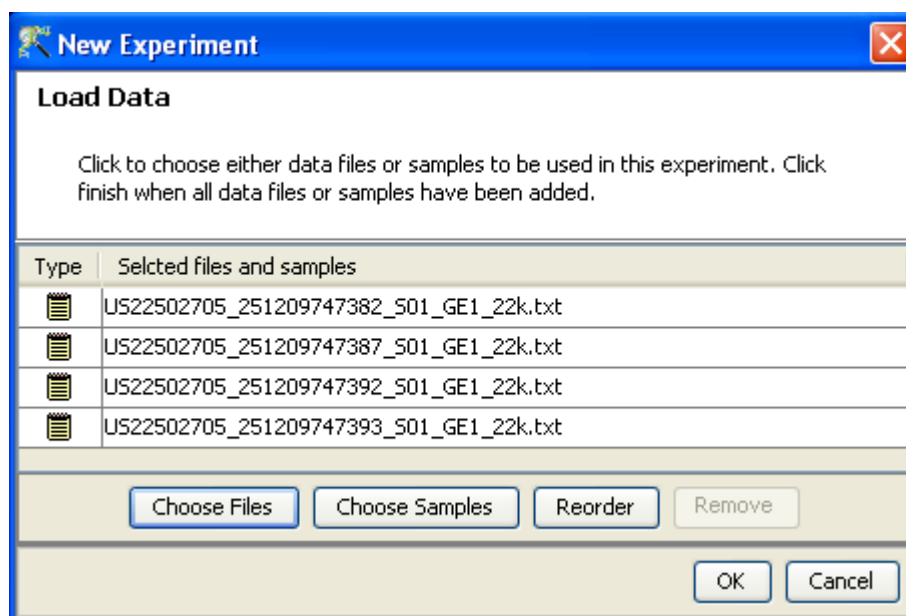


Figure 11.5: Load Data

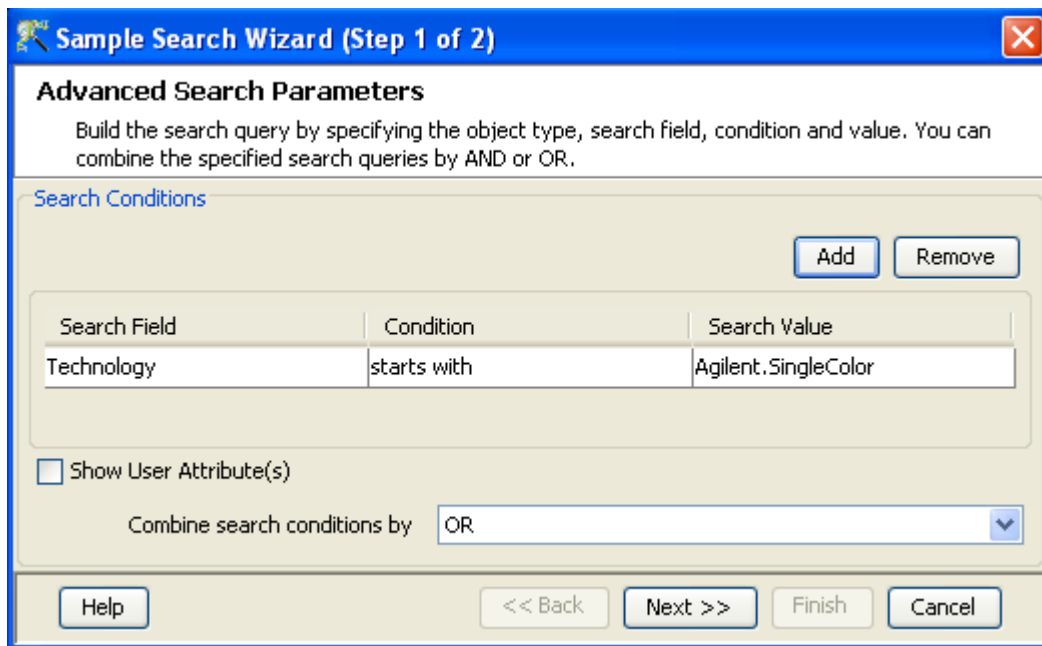


Figure 11.6: Choose Samples

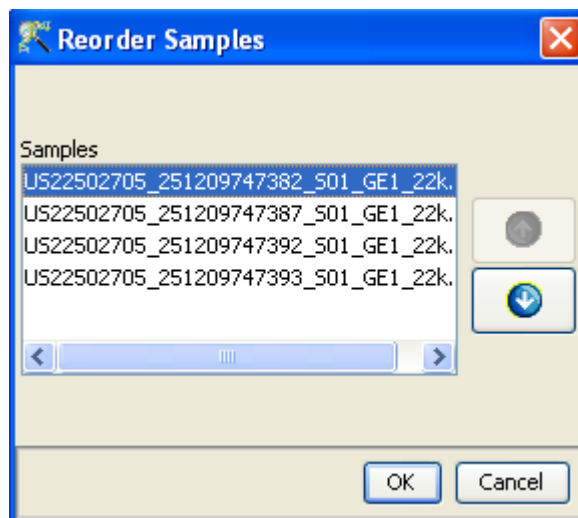


Figure 11.7: Reordering Samples



Figure 11.8: Confirmation Dialog Box

following samples of cy3/cy5: a/b, b/c, c/d, and d/a (loop design). Although you don't have sample a/c, you can still make that comparison through b. This allows you to make more comparison pairs using less chips. However, a loop design needs special handling from an analysis perspective. So, an overly simplistic approach is to split the channels and treat them as one-color data. Another experimental design where the channel-splitting can be done is cy3/cy5: a/b, c/d, e/f and g/h. Here 'a' can be compared with either b, d, f or h.

The Two Color data from Agilent FE is Lowess normalized. After the channel-splitting, it is recommended to perform either Quantile or Median Shift normalization as well.

When an Agilent Single Color experiment is created using an Agilent Two Color file as input, a message appears asking the user if a Single Color experiment needs to be created. Clicking on *OK* splits the channels and an experiment is created. This is seen in Figure 11.8 Upon clicking *OK*, the Agilent Single Color workflow appears.

The *Guided Workflow* wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

11.2 Data Processing for Agilent Single Color arrays

- **File formats:** The data files should be in text (.txt) format and obtained from Agilent Feature Extraction (FE).
- **Raw Signal Values:** The term "raw" signal values refer to the linear data after thresholding and summarization. Summarization is performed by computing the geometric mean.
- **Normalized Signal Values:** "Normalized" value is the value generated after log transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and baseline transformation.
- **Treatment of on-chip replicates:** For each replicate with multiple flags, the order of importance

	Signal	f1	f2	f3	(Resultant flag, A>M>P)
p1	1	P	M	A	A
p1	2	P	M	M	M
p1	3	P	P	P	P
p1	4	M	M	P	M
p1	5	M	P	M	M

Overall flag for p1 (Exclude A and assign majority): M
Overall Signal = (2+3+4+5)/4 = 3.5

Figure 11.9: Agilent Single Colour - Handling on chip replicates: Example 1

	Signal	f1	f2	f3	(Resultant flag, A>M>P)
p1	1	P	M	A	A
p1	2	A	M	P	A
p1	3	M	A	P	A
p1	4	A	A	P	A
p1	5	A	A	A	A

Overall flag for p1 (No P or M present, so take A): A
Overall Signal = (1+2+3+4+5)/5 = 3

Figure 11.10: Agilent Single Colour - Handling on chip replicates: Example 2

is Absent(A)>Marginal(M)>Present(P). If there is even one A, then the resultant flag is 'A'. If there is no A, but M and P, then M is assigned. If there are only Ps then only the resultant flag is assigned as 'P'. To get the overall flag for all replicates, **GeneSpring GX** excludes 'A' flag and assigns the majority considering the remaining ones. If there are only 'A' flags, only then the overall flag becomes 'A'. The following two examples illustrate this.

- **Flag values:** The flag value of a particular probeset is dependant on the flag values of the probes in it. If a probeset contains a probe which is marked as Present (P), the probeset is marked as P irrespective of the other flag values. The order of importance for flag values is Present>Marginal>Absent.
- **Treatment of Control probes:** The control probes are included while performing normalization. However there should be an exact match between the control probes in the technology and the sample for the probes to be utilized, as the comparison between the identifier columns is case-sensitive.
- **Empty Cells:** Not Applicable.
- **Sequence of events:** The sequence of events involved in the processing of the data files is: Thresholding → Summarization (summarization is performed by computing the geometric mean) → log transformation → Normalization → Baseline Transformation.

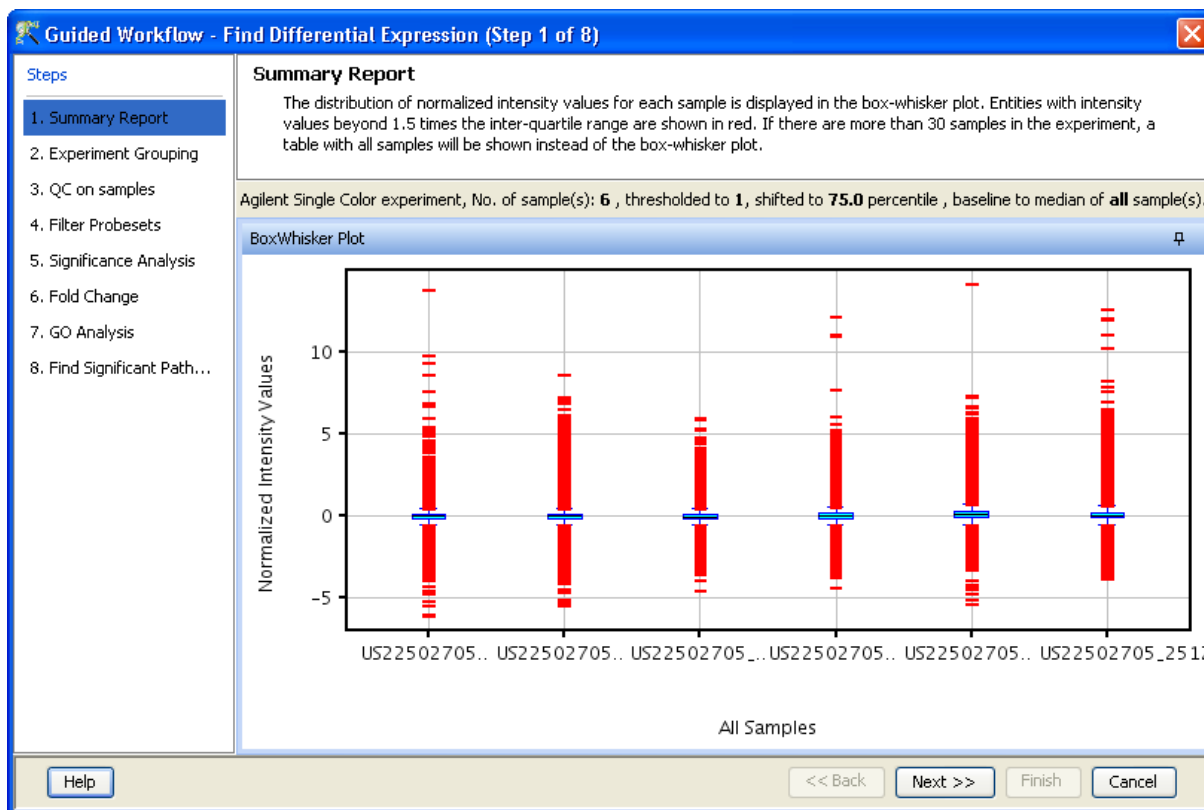


Figure 11.11: Summary Report

11.3 Guided Workflow steps



Summary report (Step 1 of 8): The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized Expression values on the Y axis. An information message on the top of the wizard shows the number of samples in the file and the sample processing details. By default, the *Guided Workflow* does a thresholding of the signal values to 5. It then normalizes the data to 75th percentile and performs baseline transformation to median of all samples. If the number of samples are more than 30, they are only represented in a tabular column. On clicking the *Next* button it will proceed to the next step and on clicking *Finish*, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. On doing a right click, the options of invert selection is displayed and on clicking the same the selection is inverted i.e., all the probes except the selected ones are highlighted in green. Figure 11.11 shows the Summary report with box-whisker plot.

Note: In the *Guided Workflow*, these default parameters cannot be changed. To choose different parameters use *Advanced Analysis*.

Experiment Grouping (Step 2 of 8): On clicking *Next*, the *Experiment Grouping* window appears which is the 2nd step in the **Guided Workflow**. It requires parameter values to be defined to

group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the **Add parameter** button. Parameter values can be assigned by first selecting the desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on **Clear**. Press **OK** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Guided Workflow**. The other parameters can be used in the **Advanced Analysis**.





Note: The *Guided Workflow* does not proceed further without grouping information.

Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file  icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

```
A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50
```

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file  icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click → *Properties* → *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click → *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

Note: The *Guided Workflow* by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the guided wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures [11.12](#) and [11.13](#) respectively.

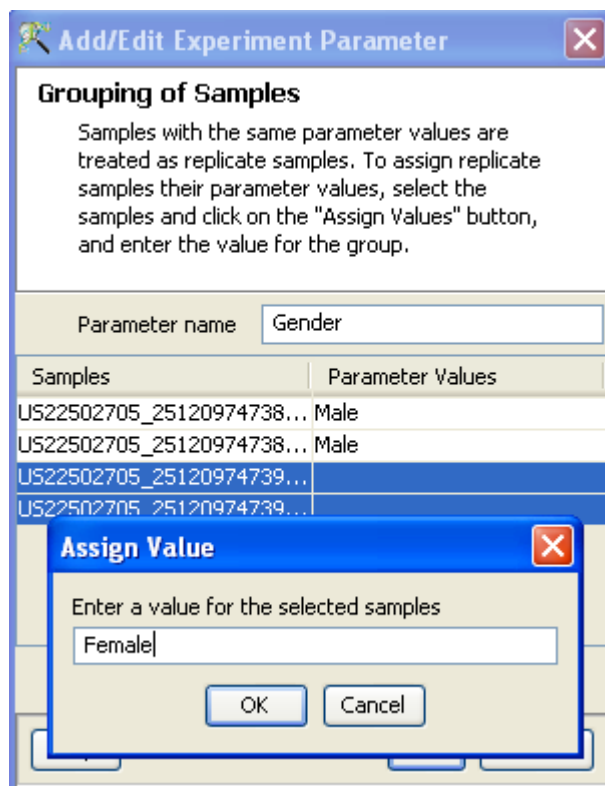


Figure 11.12: Experiment Grouping

Quality Control (Step 3 of 8): The 3rd step in the Guided workflow is the QC on samples which is displayed in the form of four tiled windows. They are as follows:

- Quality controls Metrics- Report and Experiment grouping tabs
- Quality Controls Metrics- Plot
- 3D PCA scores.
- Legend

QC on Samples generates four tiled windows as seen in Figure 11.14.

The *Metrics Report* has statistical results to help you evaluate the reproducibility and reliability of your single color microarray data.

The table shows the following:

More details on this can be obtained from the Agilent Feature Extraction Software Reference Guide, available from http://www.chem.agilent.com/Library/usermanuals/Public/G4460-90017_FE.10.5.Installation.pdf

Quality controls *Metrics Plot* shows the QC metrics present in the QC report in the form of a plot. *Principal Component Analysis (PCA)* calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster

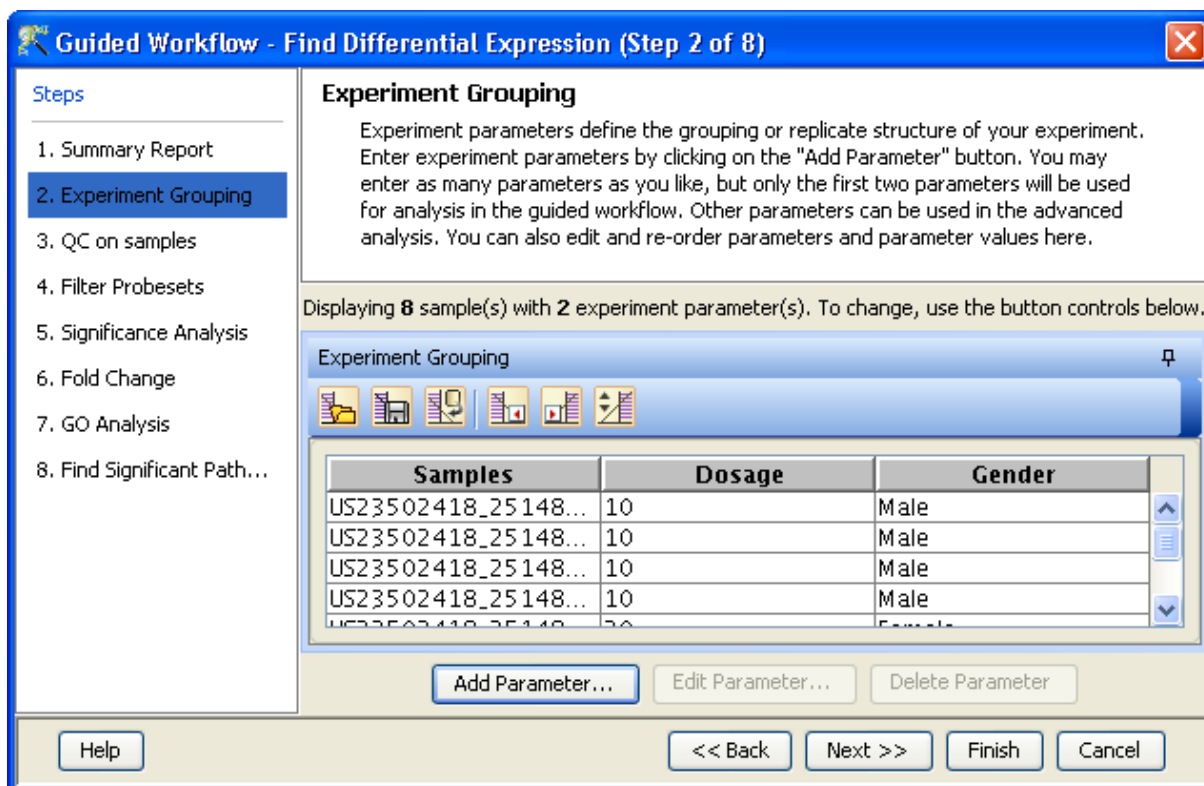


Figure 11.13: Edit or Delete of Parameters

together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, normalization as well as baseline transformation is performed again on the samples. Click on *OK* to proceed.

The fourth window shows the legend of the active QC tab.

Filter probesets (Step 4 of 8): In this step, the entities are filtered based on their flag values *P(present)*, *M(marginal)* and *A(absent)*. Only entities having the present and marginal flags in at least 1 sample are displayed in the profile plot. The selection can be changed using *Rerun Filter* option. The flagging information is derived from the Feature columns in data file. More details on how flag values [P,M,A] are calculated can be obtained from <http://www.chem.agilent.com>. The plot is generated using the normalized signal values and samples grouped by the active interpretation. Options to customize the plot can be accessed via the Right-click menu. An *Entity List*, corresponding to this filtered list, will be generated and saved in the Navigator window. The Navigator window can be viewed after exiting from *Guided Workflow*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Newer annotations can be added and existing ones removed using the *Configure Columns* button. Additional tabs in

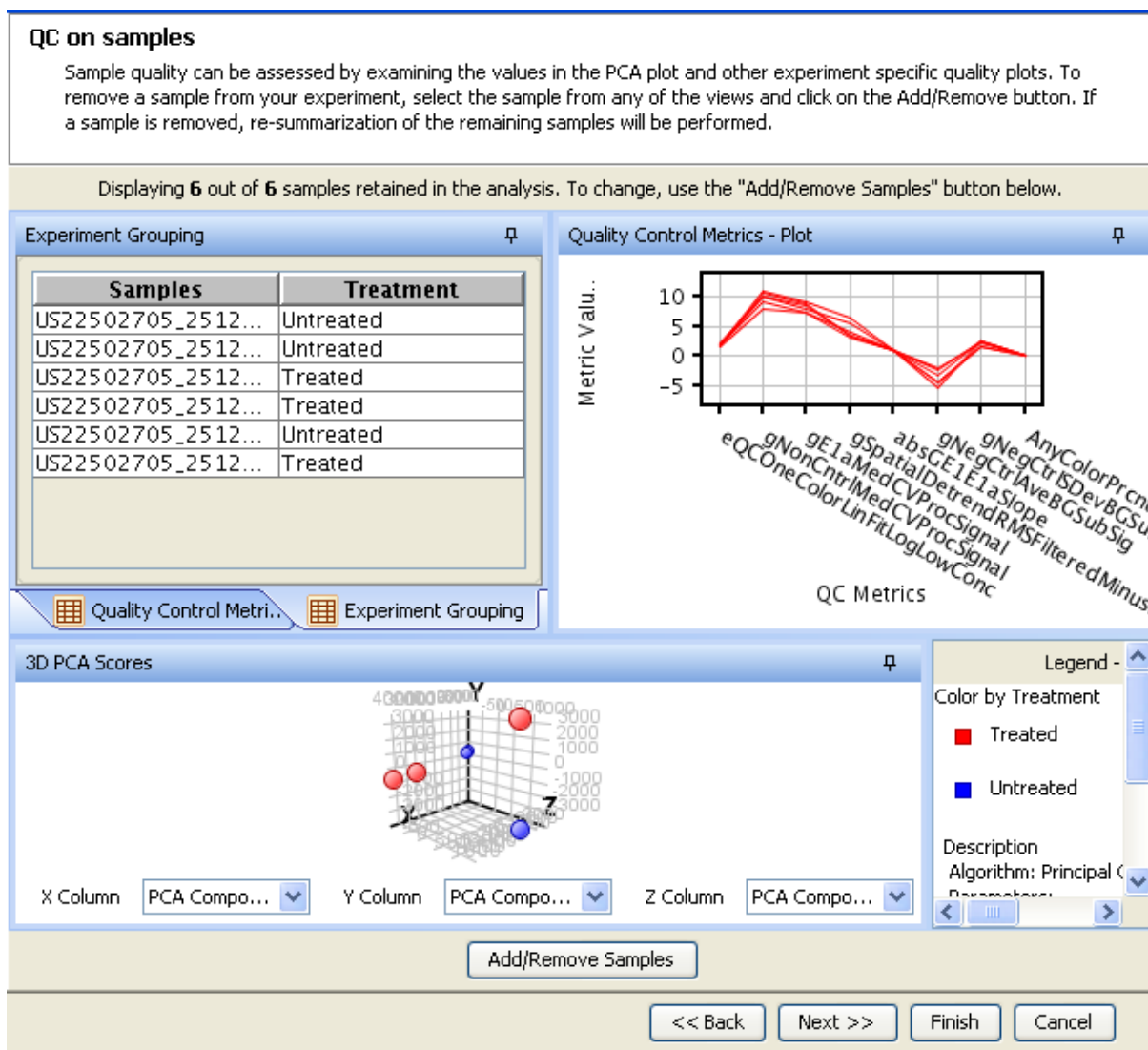


Figure 11.14: Quality Control on Samples

the *Entity Inspector* give the raw and the normalized values for that entity. The cutoff for filtering can be changed using the *Rerun Filter* button. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. The information message on the top shows the number of entities satisfying the flag values. Figures 11.15 and 11.16 are displaying the profile plot obtained in situations having single and two parameters.

Significance Analysis(Step 5 of 8): Depending upon the experimental grouping, **GeneSpring GX** performs either T-test or ANOVA. The tables below describe broadly the type of statistical test performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I*, has 2 groups, the normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.
- **Example Sample Grouping II:** In this example, only one group, the tumor, is present. T-test

Name of Metric	FE Stats Used	Description/Measures
eQCOneColor LinFitLogLowConc	eQCOneColor LinFitLogLowConc	Log of lowest detectable concentration from fit of Signal vs. Concentration of E1a probes
AnyColorPrcent BGNonUnifOL	AnyColorPrcent BGNonUnifOL	Percentage of LocalBkgdRegions that are NonUnifOlr in either channel
gNonCtrlMedCVPProcSignal	gMedPrcentCVPProcSignal	The median percent CV for replicate non-control probes using the processed signal.
gE1aMedCVPProcSignal	geQCMedPrcentCVPProcSignal	This is the same as MedPrcentCVPProcSignal, except that it is performed using the eQC SpikeIn Replicates rather than the nonControl Replicates. There must be at least 3 CVs from which to calculate a median.
gSpatialDetrend RMSFilteredMinusFit	gSpatialDetrend RMSFilteredMinusFit	Residual of background detrending fit
absGE1E1aSlope	Abs(eQCOneColor LinFitSlope)	Absolute of slope of fit for Signal vs. Concentration of E1a probes
gNegCtrl AveBGSubSig	gNegCtrl AveBGSubSig	Avg of NegControl Bkgd-subtracted signals (Green)
gNegCtrl SDevBGSubSig	gNegCtrl SDevBGSubSig	StDev of NegControl Bkgd-subtracted signals (Green)
AnyColor PrcentFeatNonUnifOL	AnyColor PrcentFeatNonUnifOL	Percentage of Features that are NonUnifOlr

Table 11.1: Quality Controls Metrics

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 11.2: Sample Grouping and Significance Tests I

against zero will be performed here.

- **Example Sample Grouping III:** When 3 groups are present (normal, tumor1 and tumor2) and one of the groups (tumor2 in this case) does not have replicates, statistical analysis cannot be performed. However if the condition tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.
- **Example Sample Grouping IV:** When there are 3 groups within an interpretation, One-way

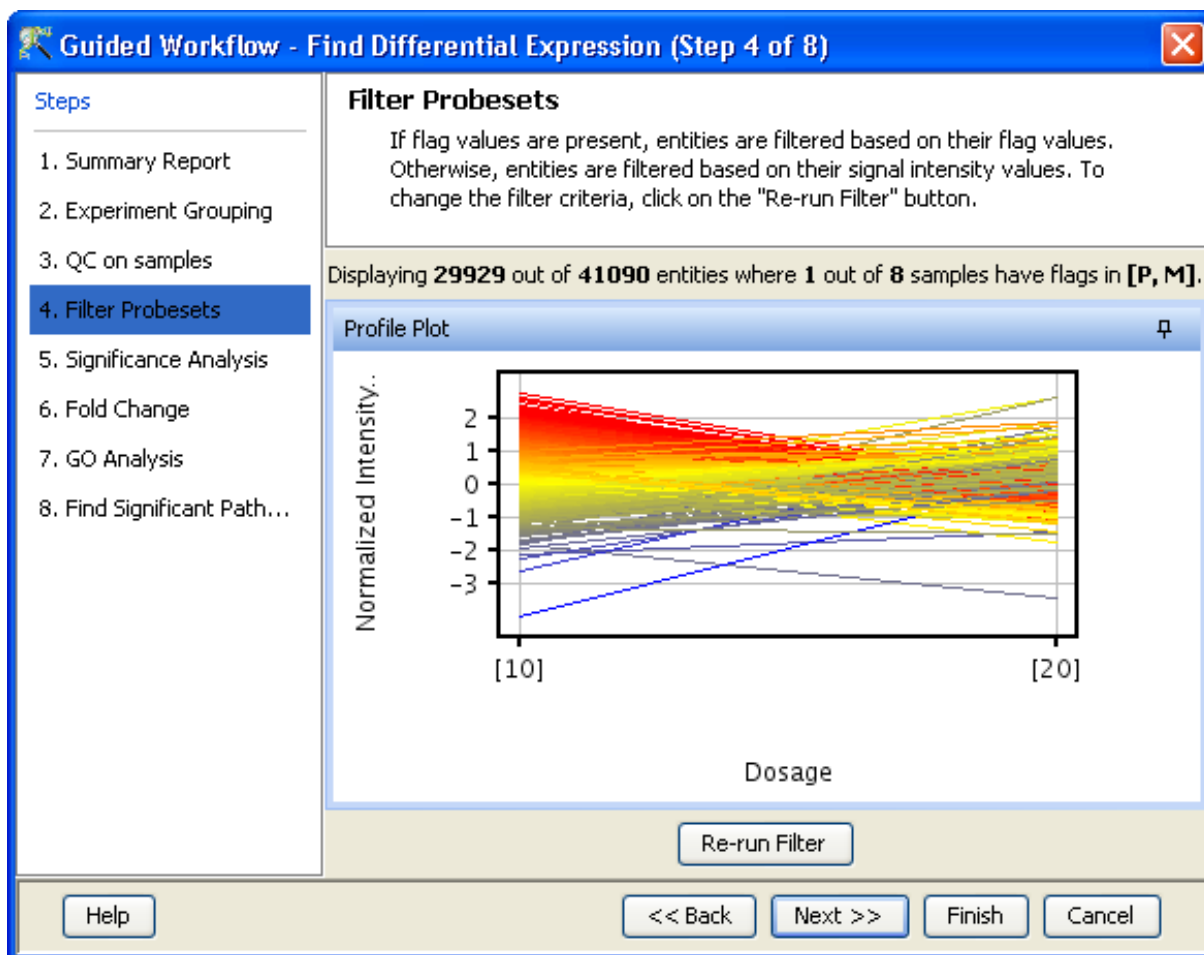


Figure 11.15: Filter Probesets-Single Parameter

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 11.3: Sample Grouping and Significance Tests II

ANOVA will be performed.

- **Example Sample Grouping V:** This table shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.
- **Example Sample Grouping VI:** In this table, a two-way ANOVA will be performed.
- **Example Sample Grouping VII:** In the example below, a two-way ANOVA will be performed

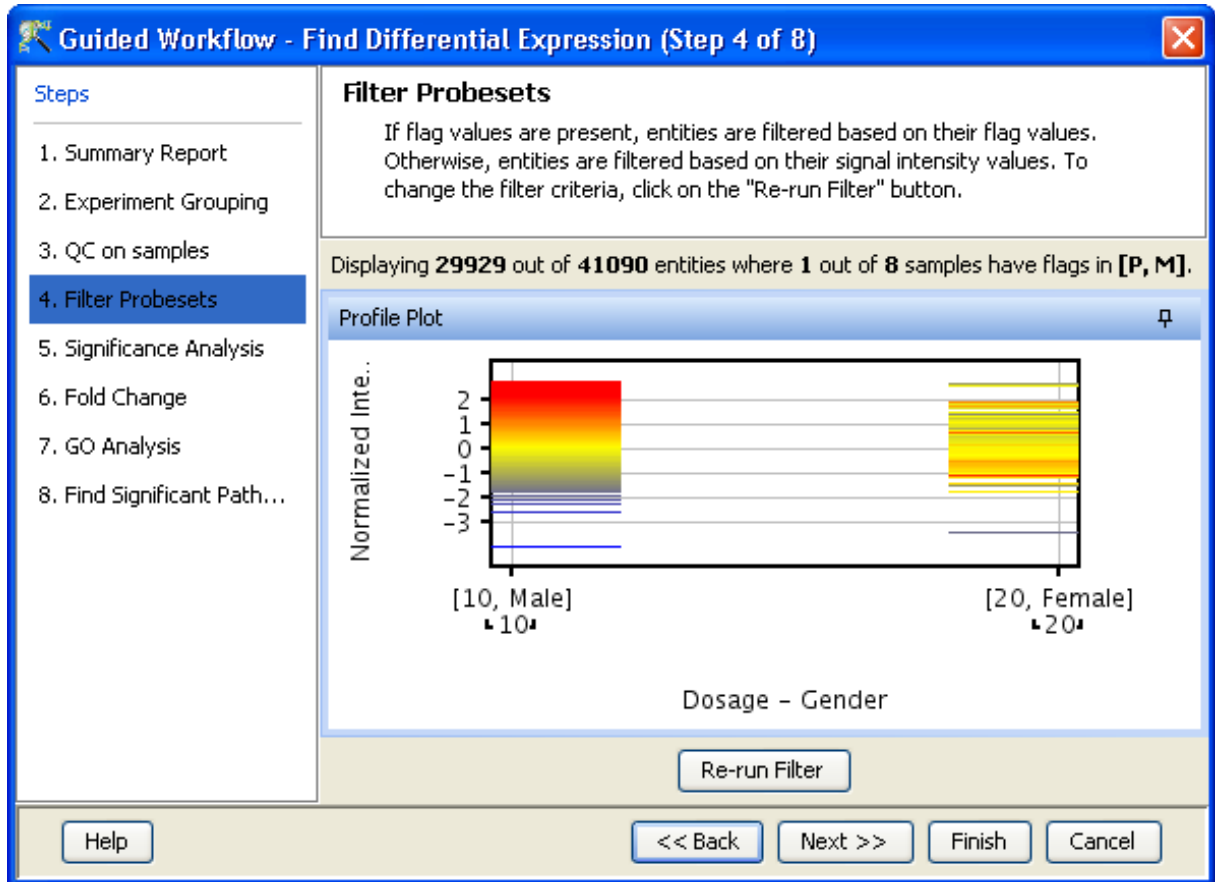


Figure 11.16: Filter Probesets-Two Parameters



Figure 11.17: Rerun Filter

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 11.4: Sample Grouping and Significance Tests III

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 11.5: Sample Grouping and Significance Tests IV

and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Statistical Tests: T-test and ANOVA

- **T-test: T-test unpaired** is chosen as a test of choice with a kind of experimental grouping shown in Table 1. Upon completion of T-test the results are displayed as three tiled windows.
 - A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *Regulation*.
 - *Differential expression analysis report* mentioning the Test description i.e. test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic or Permutative*).

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

- **Analysis of variance(ANOVA)**: ANOVA is chosen as a test of choice under the experimental grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:
- A *p-value table* consisting of probe names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 11.6: Sample Grouping and Significance Tests V

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 11.7: Sample Grouping and Significance Tests VI

- *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and p-value computation type (*Asymptotic or Permutative*).
- *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

Special case: In situations when samples are not associated with at least one possible permutation of conditions (like Normal at 50 min and Tumor at 10 min mentioned above), no p-value can be computed and the **Guided Workflow** directly proceeds to **GO analysis**.

Fold-change (Step 6 of 8): Fold change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis. The wizard shows a table consisting of 3 columns: Probe Names, Fold change value and regulation (up or down). The regulation column depicts which one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using *Re-run Filter*. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1. A profile plot is also generated. Upregulated entities are shown in red. The color can be changed using the Right-click→*Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cutoff in the experiment Navigator.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 11.8: Sample Grouping and Significance Tests VII

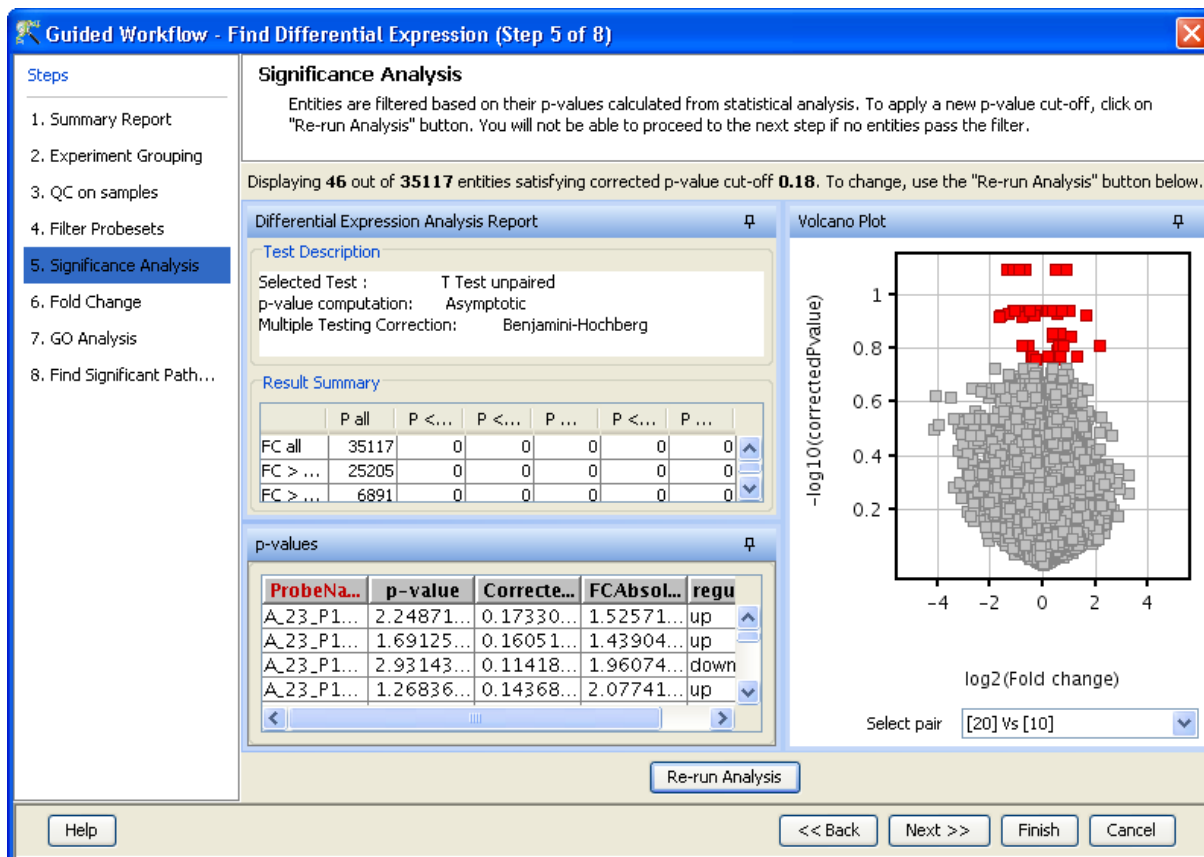


Figure 11.18: Significance Analysis-T Test

Note: Fold Change step is skipped and the *Guided Workflow* proceeds to the *GO Analysis* in case of experiments having 2 parameters.

Fold Change view with the spreadsheet and the profile plot is shown in Figure 11.20.

Gene Ontology Analysis(Step 7 of 8): The *GO Consortium* maintains a database of controlled vocabularies for the description of molecular function, biological process and cellular location of gene products. The GO terms are displayed in the Gene Ontology column with associated *Gene Ontology Accession* numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more

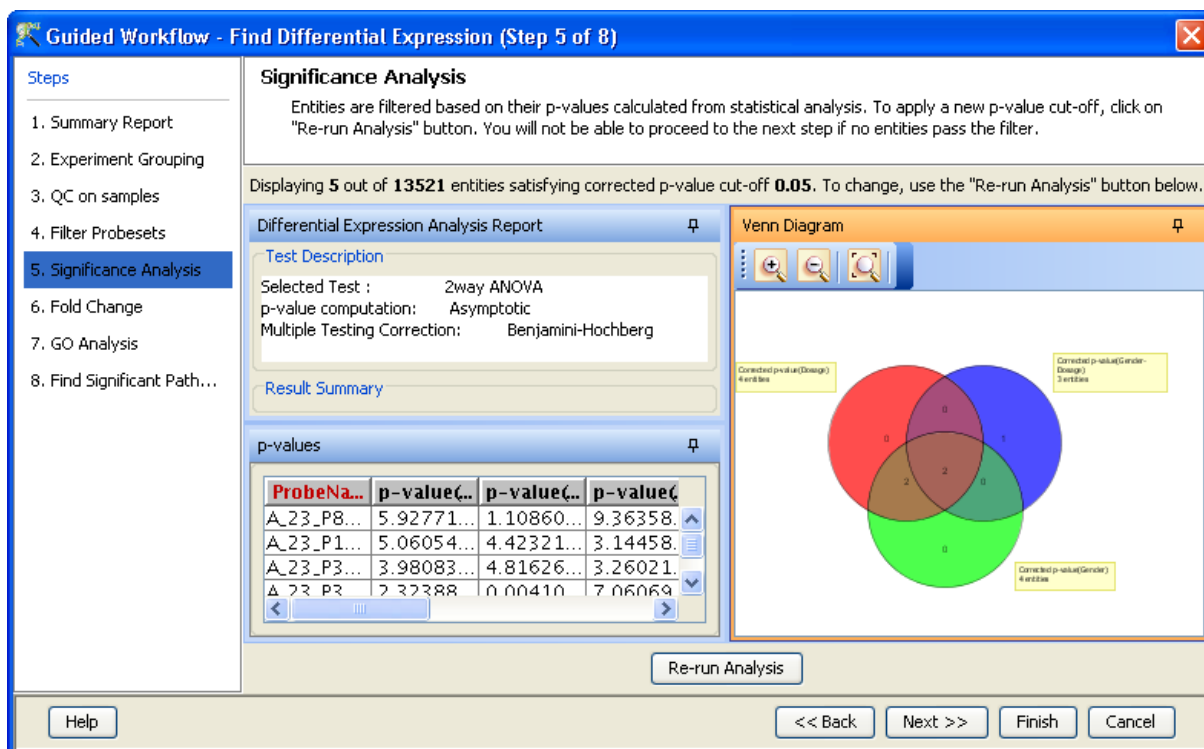


Figure 11.19: Significance Analysis-Anova

parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

The GO analysis wizard shows two tabs comprising of a spreadsheet and a *GO tree*. The *GO Spreadsheet* shows the *GO Accession* and *GO terms* of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However, the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.1 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while

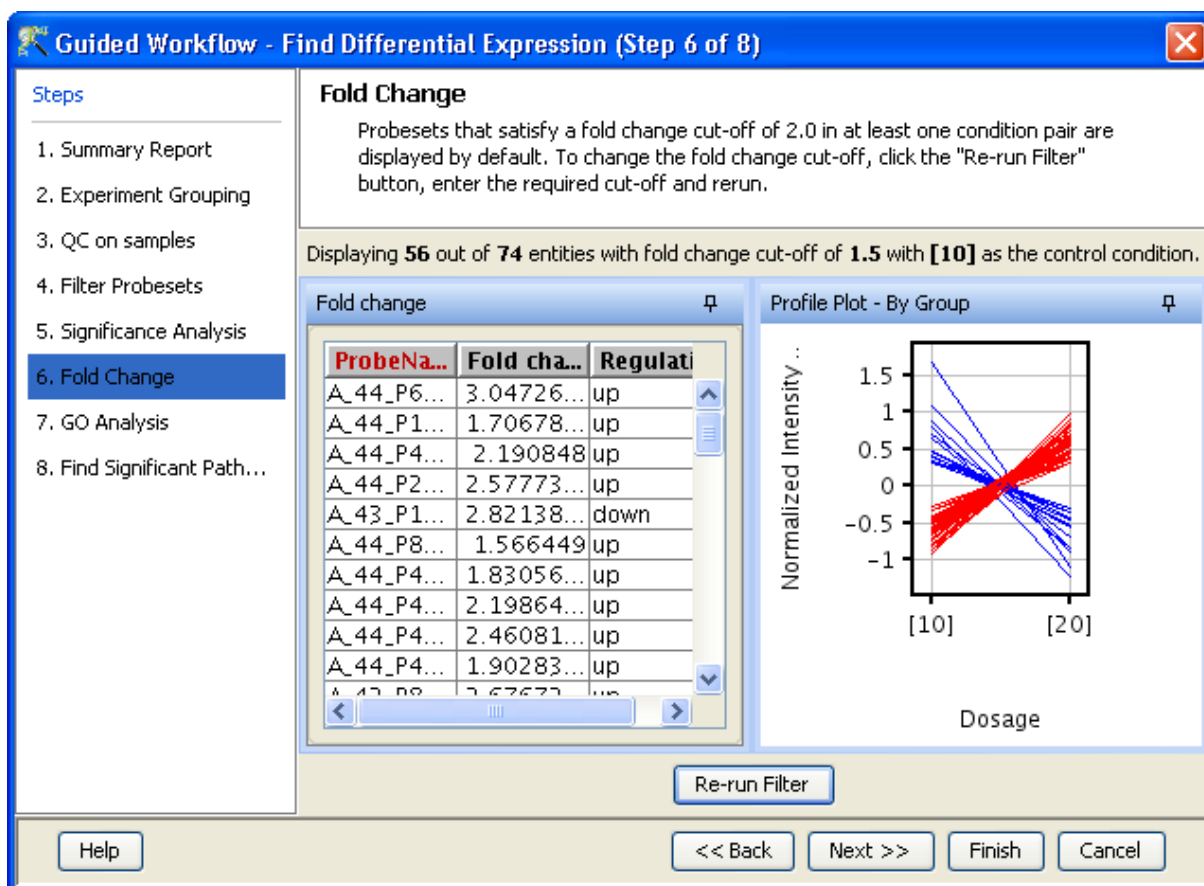


Figure 11.20: Fold Change

others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button is also provided on the GO tree panel to search using some keywords

Note : In **GeneSpring GX** GO analysis implementation, all the three component: Molecular Function, Biological Processes and Cellular location are considered together.

On finishing the GO analysis, the *Advanced Workflow* view appears and further analysis can be carried out by the user. At any step in the Guided workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any) and the *Advanced Workflow* view appears.

Find Significant Pathways (Step 8 of 8): This step in the Guided Workflow finds relevant pathways from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the Fold Change (step 6 of 8). This view shows two tables-

- The Significant Pathways table shows the names of the pathways as well as the number of nodes and entities in the pathway and the p-values. It also shows the number of entities that are

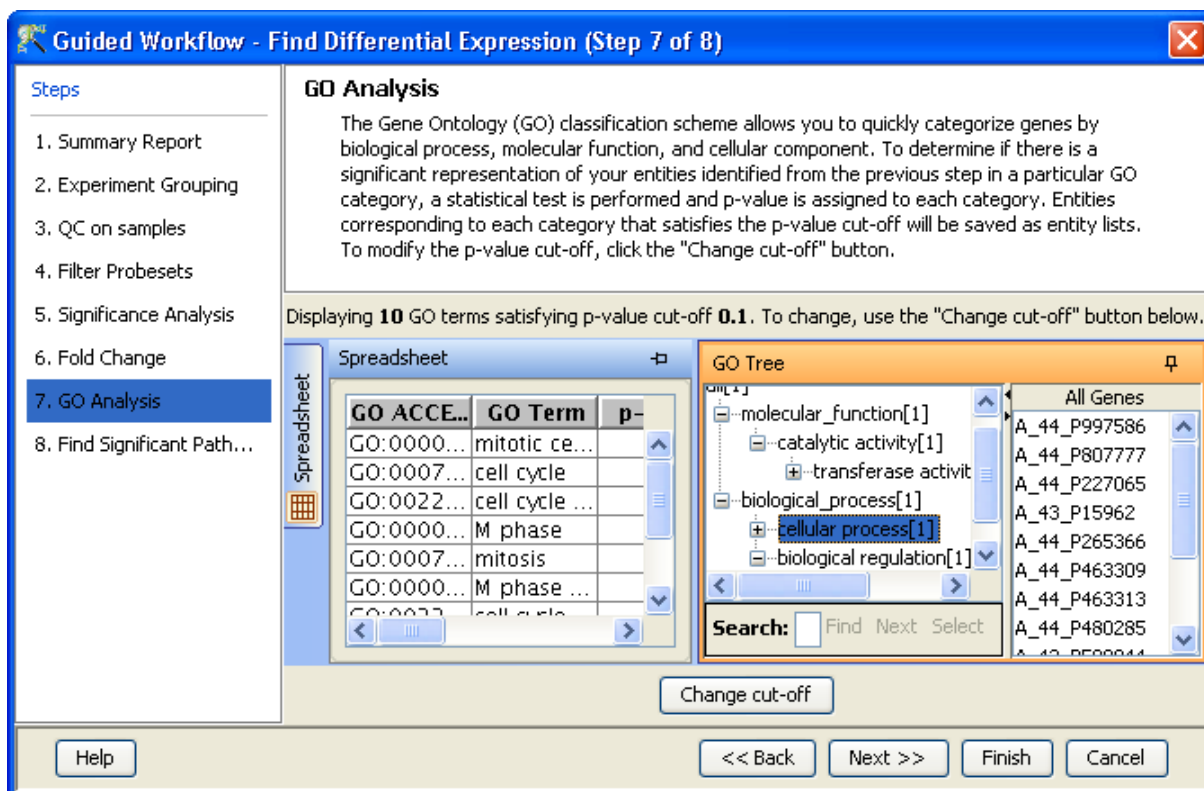


Figure 11.21: GO Analysis

similar to the pathway and the entity list. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.

- The Non-significant Pathways table shows the pathways in the tool that do not have a single entity in common with the ones in the given entity list.

The user has an option of changing the p-value cut-off(using *Change cutoff*) and also to save specific pathways using the *Custom Save* option. See figure 11.22. On clicking, *Finish* the main tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of the Guided Workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in the Guided Workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any).

Note: In case the user is using **GeneSpring GX** for the first time, this option will give results using the demo pathways. The user can upload the pathways of his/her choice by using the option *Import BioPAX pathways* under **Tools** in the **Menu** bar. Later instead of reverting to the Guided Workflow the user can use the option *Find Significant Pathways* in **Results Interpretation** under the same Workflow.

The default parameters used in the *Guided Workflow* is summarized below

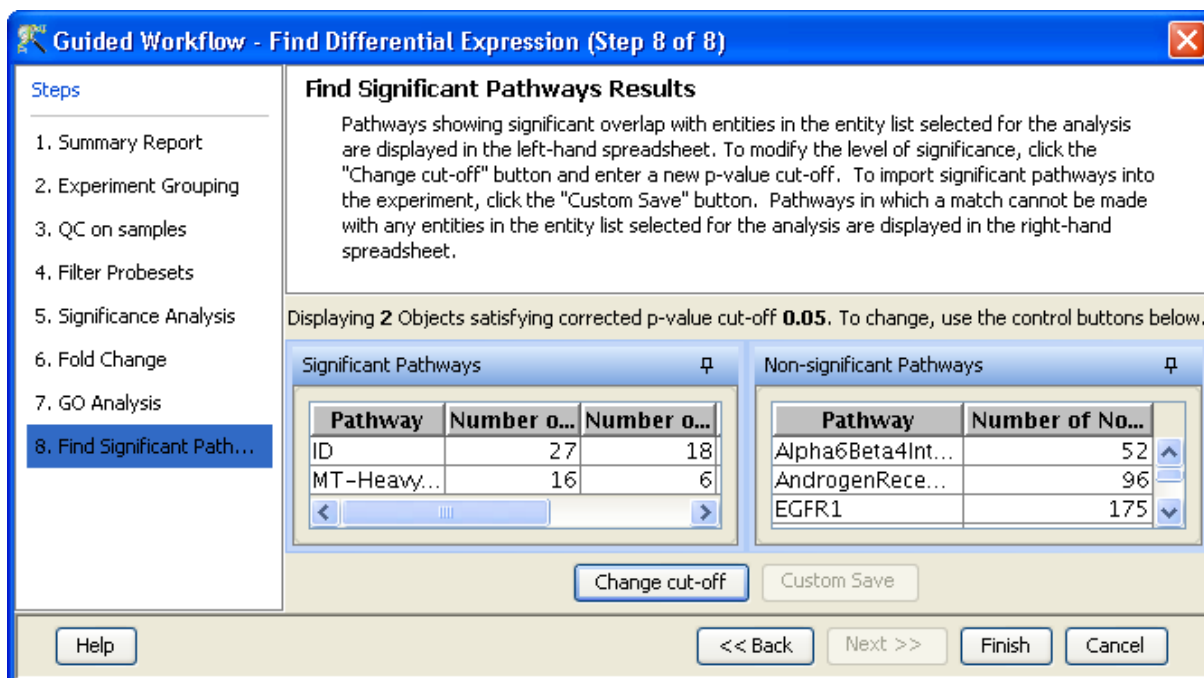


Figure 11.22: Find Significant Pathways

	Parameters	Parameter values
Expression Data Transformation	Thresholding	1.0
	Normalization	Shift to 75th Percentile
	Baseline Transformation	Median to all samples
	Summarization	Not Applicable
Filter by		
1.Flags	Flags Retained	Present(P), Marginal(M)
2.Expression Values	(i) Upper Percentile cutoff	Not Applicable
	(ii) Lower Percentile cutoff	
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1
Find Significant Pathways	p-value cutoff	0.05

Table 11.9: Table of Default parameters for Guided Workflow

11.4 Advanced Workflow

The *Advanced Workflow* offers a variety of choices to the user for the analysis. Flag options can be changed and raw signal thresholding can be altered. Additionally there are options for baseline transformation of

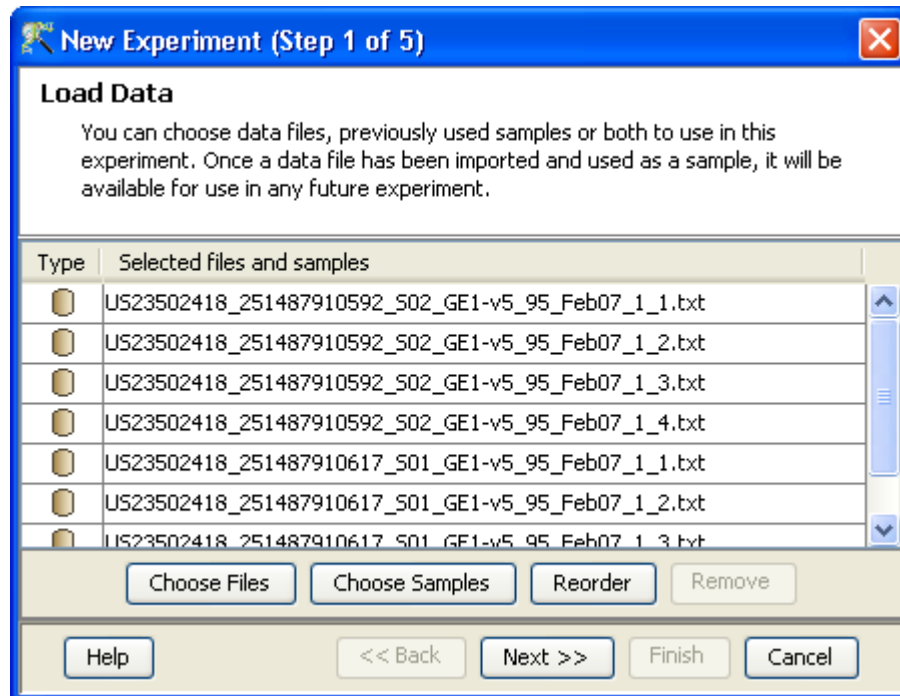


Figure 11.23: Load Data

the data and for creating different interpretations. To create and analyze an experiment using the *Advanced Workflow*, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Advanced. Clicking on **OK** will open a new experiment wizard which then proceeds as follows:

1. **Step 1 of 5: Load Data** As in case of *Guided Workflow*, either data files can be imported or else pre-created samples can be used.
 - For loading new txt files, use *Choose Files*.
 - If the txt files have been previously used in **GeneSpring GX** experiments *Choose Samples* can be used.

Step 1 of Experiment Creation, the 'Load Data' window, is shown in Figure 11.23.

2. **Step 2 of 5: Advanced Flag Import** This gives the options for importing flag information. The information is derived from the Feature columns in data file. User has the option of changing the default flag settings that appear in this step. The 'Save as Default' handle allows saving the current flag settings under the tool configuration. When a file is imported, **GeneSpring GX** will show these saved default setting in this step, by default. The settings can be changed either in this wizard or from *Tools* → *Options* → *Miscellaneous* → *Agilent Flag Settings*.

Step 2 of Experiment Creation, the 'Advanced flag Import' window, is depicted in the Figure 11.24.

3. **Step 3 of 5: Normalization Options**

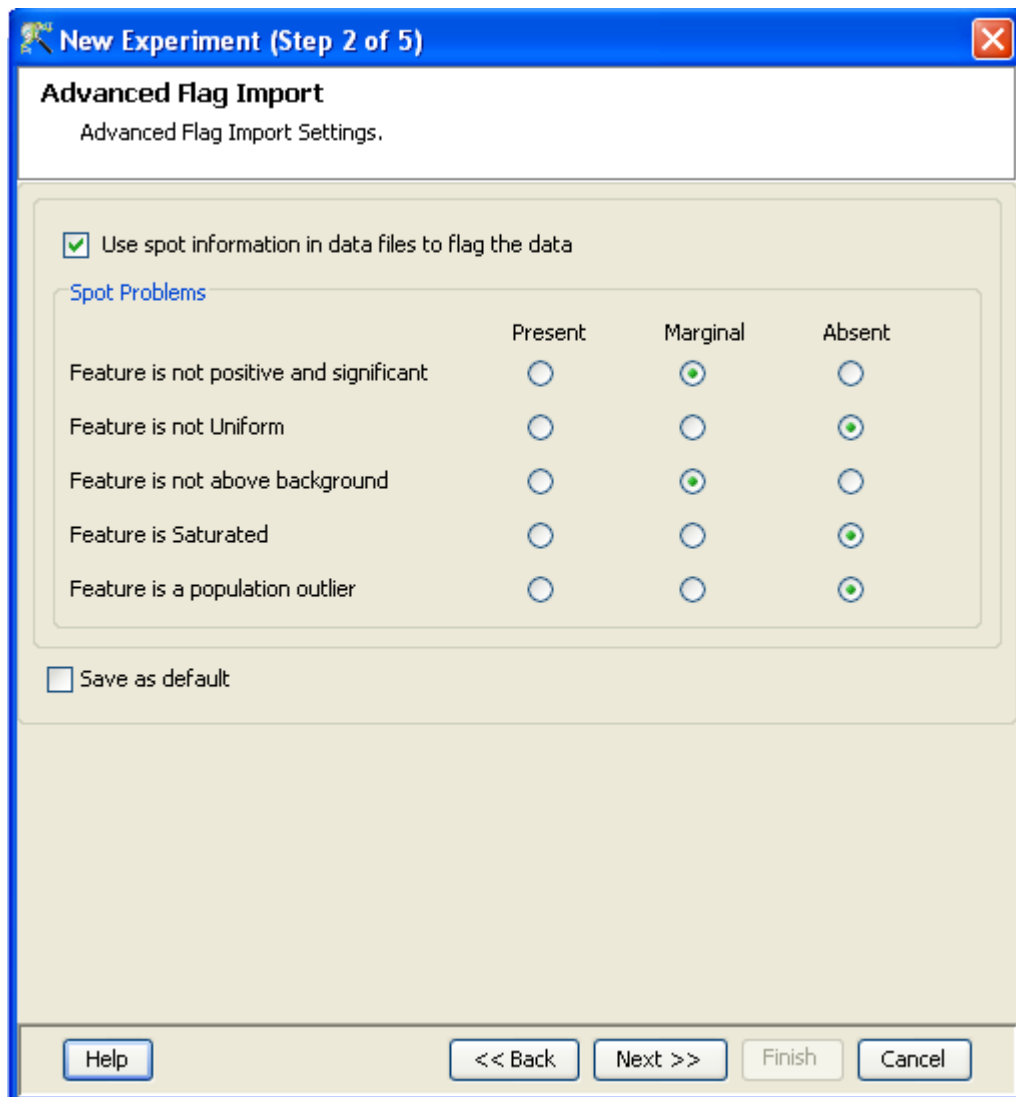


Figure 11.24: Advanced flag Import

Criteria for preprocessing of input data is set here. It allows the user to threshold raw signals to chosen values and select normalization algorithms (None, Percentile Shift, Scale, Quantile, Normalize to control genes or Normalize to External Value).

- **None:** No normalization is done.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

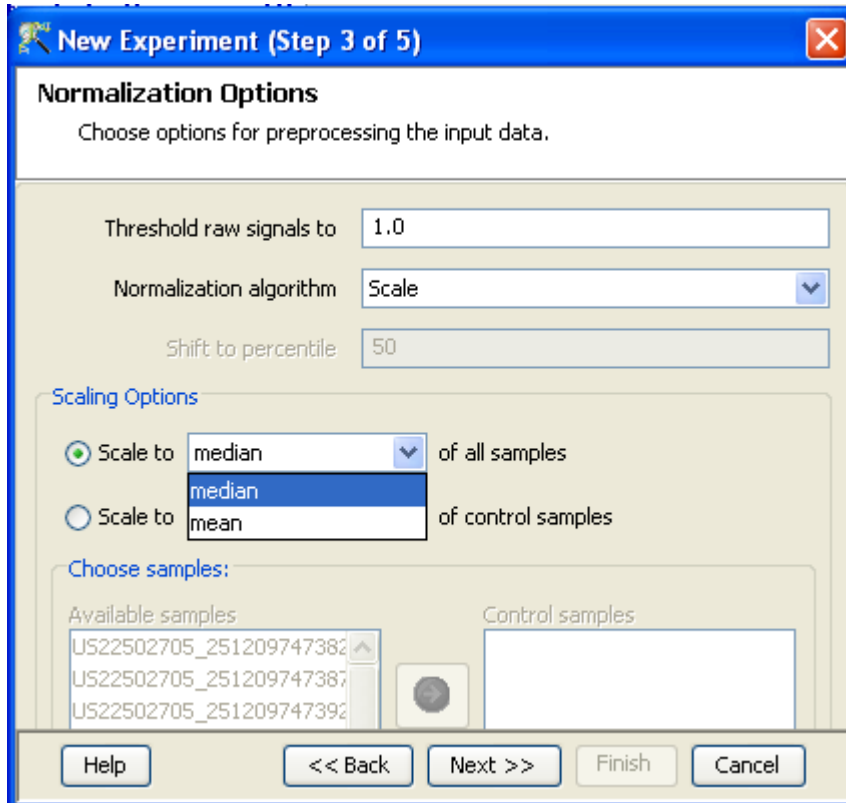


Figure 11.25: Preprocess Options

- **Quantile:** Makes all statistical parameters of the sample, ie, mean, median and percentile identical.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the '*Assign Value*' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.

For details on the above normalization methods, refer to section [Normalization Algorithms](#).

Figure 11.25 shows the Step 3 of Experiment Creation.

Step 5: Choose entities If the **Normalize to control genes** option was chosen in step 3, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user

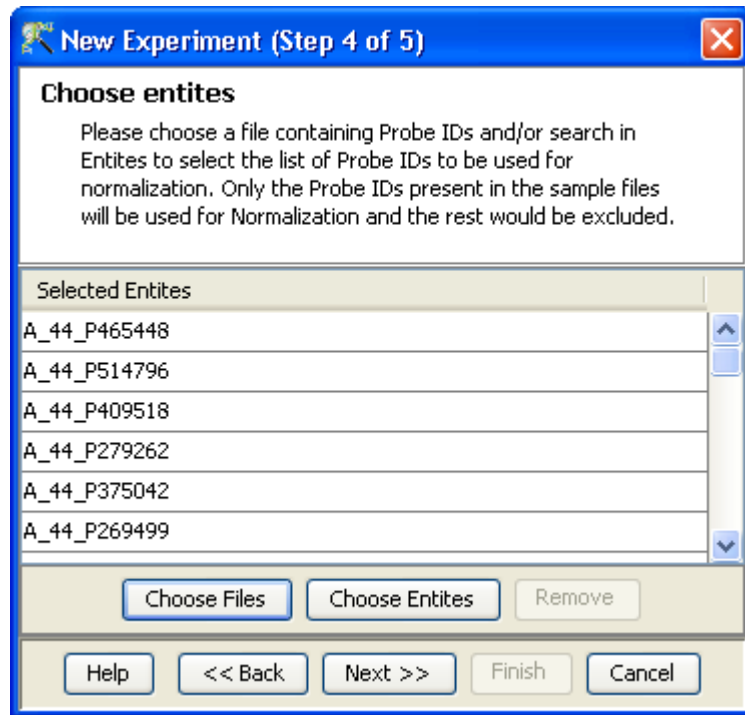


Figure 11.26: Normalize to control genes

has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same. See figure 11.26.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

ss Baseline Options This step allows the user to perform baseline transformation. See figure 11.27.

The baseline options include:

- *Do not perform baseline* No transformation is done.
- *Baseline to median of all samples:* For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- *Baseline to median of control samples:* For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on *Assign value*. This opens up the *Choose Control Samples* window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on *Ok*. This will show the control samples for each of the samples.

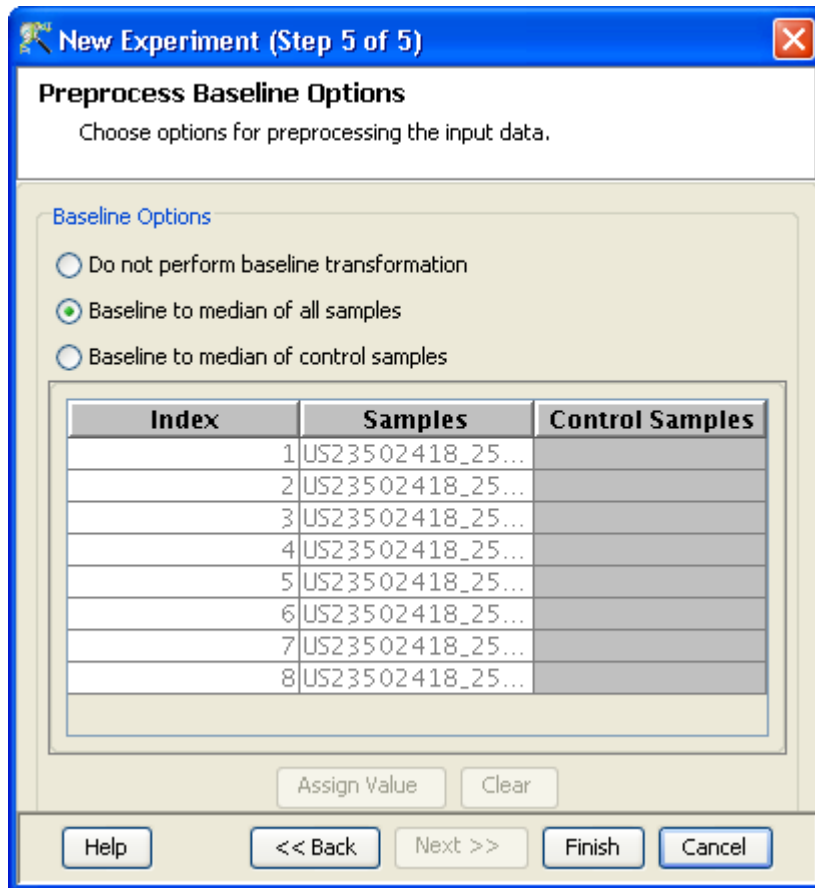


Figure 11.27: Baseline Transformation Options

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to *View* in Toolbar.

Once an experiment is created, the *Advanced Workflow* steps appear on the right hand side. Following is an explanation of the various workflow links:

11.4.1 Experiment Setup

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the Advanced Workflow, the method of analysis, the details of the algorithms used and the interpretation of results

- **Experiment Grouping:** *Experiment Parameters* defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

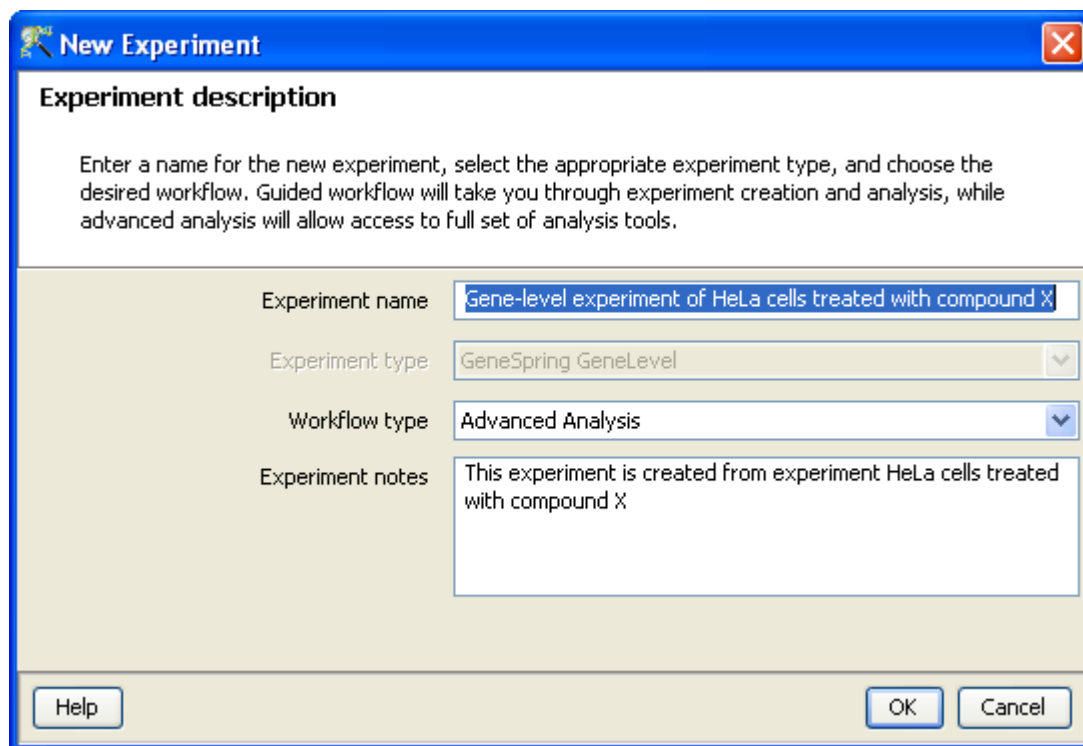


Figure 11.28: Gene Level Experiment Creation

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear

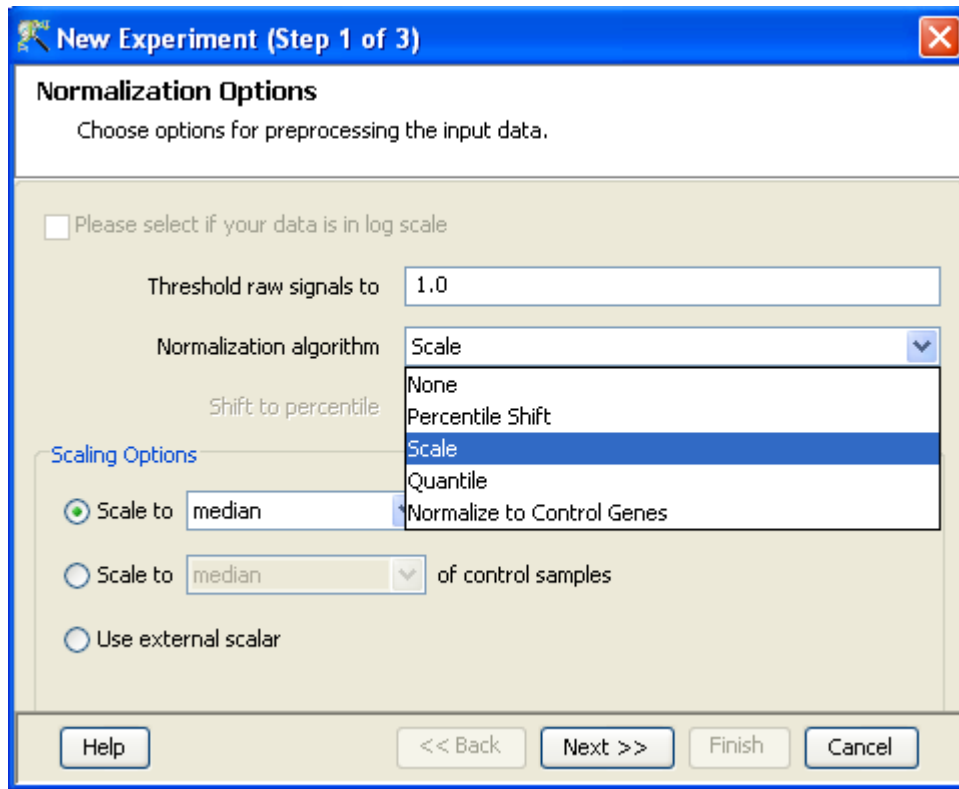


Figure 11.29: Gene Level Experiment Creation - Normalization Options

in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

11.4.2 Quality Control

- **Quality Control on Samples:**

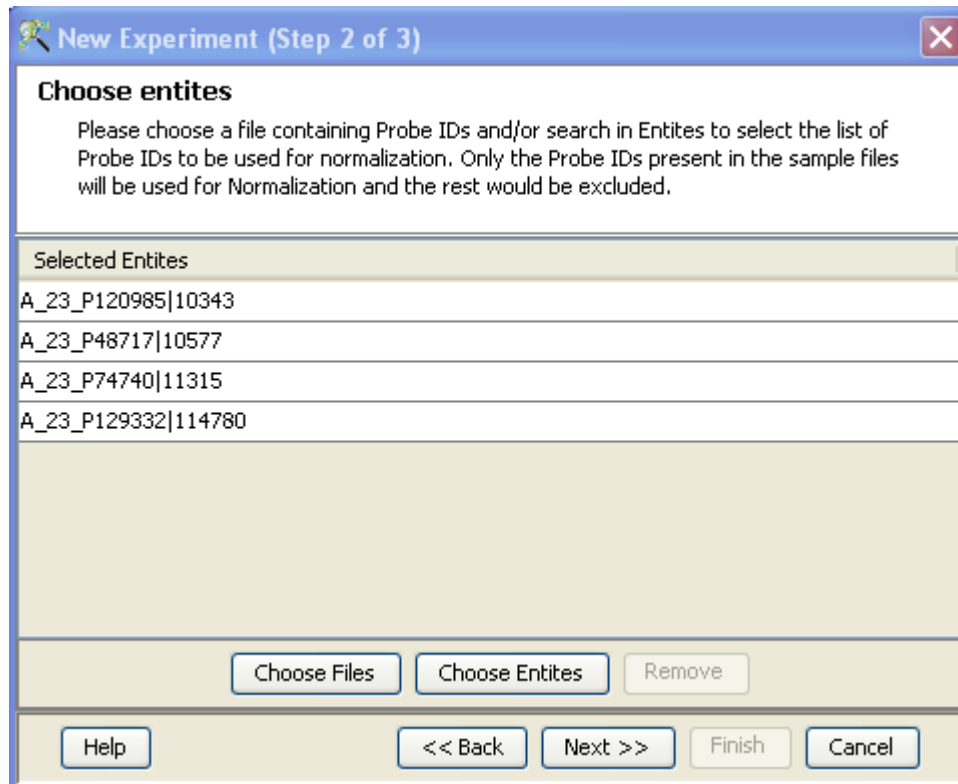


Figure 11.30: Gene Level Experiment Creation - Choose Entities

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows four tiled windows:

- Correlation plots and Correlation coefficients
- Quality Metrics Report and Quality Metrics plot and experiment grouping tabs.
- PCA scores
- Legend

Figure 11.32 has the 4 tiled windows which reflect the QC on samples.

The *Correlation Plots* shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$

The heatmap is colorable by Experiment Factor information via Right-Click → Properties. Similarly, the intensity levels in the heatmap are also customizable.

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

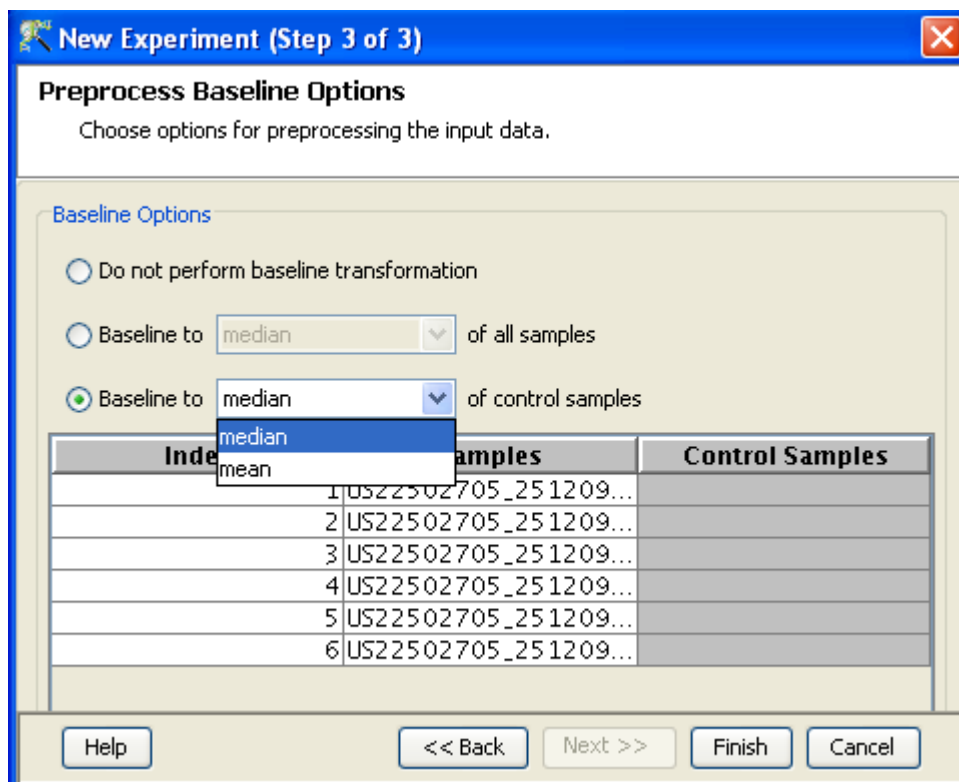


Figure 11.31: Gene Level Experiment Creation - Preprocess Baseline Options

The metrics report include statistical results to help you evaluate the reproducibility and reliability of your single microarray data.

More details on this can be obtained from the Agilent Feature Extraction Software(v9.5) Reference Guide, available from <http://chem.agilent.com>.

Quality controls Metrics Plot shows the QC metrics present in the QC report in the form of a plot.

Experiment Grouping shows the parameters and parameter values for each sample.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The fourth window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on *OK* to proceed.

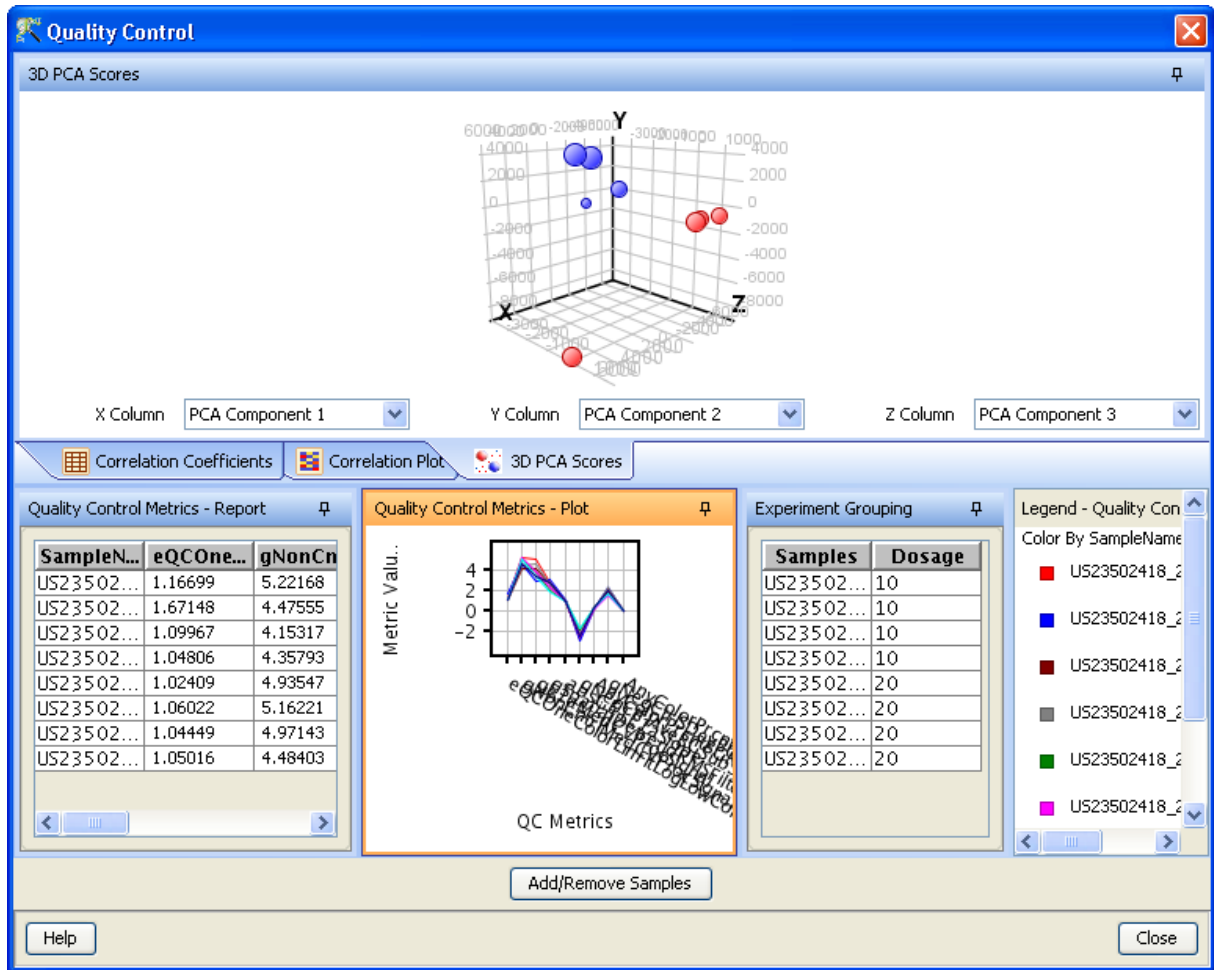


Figure 11.32: Quality Control

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags:** In this step, the entities are filtered based on their flag values, the P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new experiment (Step 2 of 3) are taken into consideration while filtering the entities. The filtration is done in 4 steps:
 1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window. This is seen in [Figure 11.33](#)
 2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box (See [Figure 11.34](#)).
 3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window (See [Figure 11.35](#))



Figure 11.33: Entity list and Interpretation

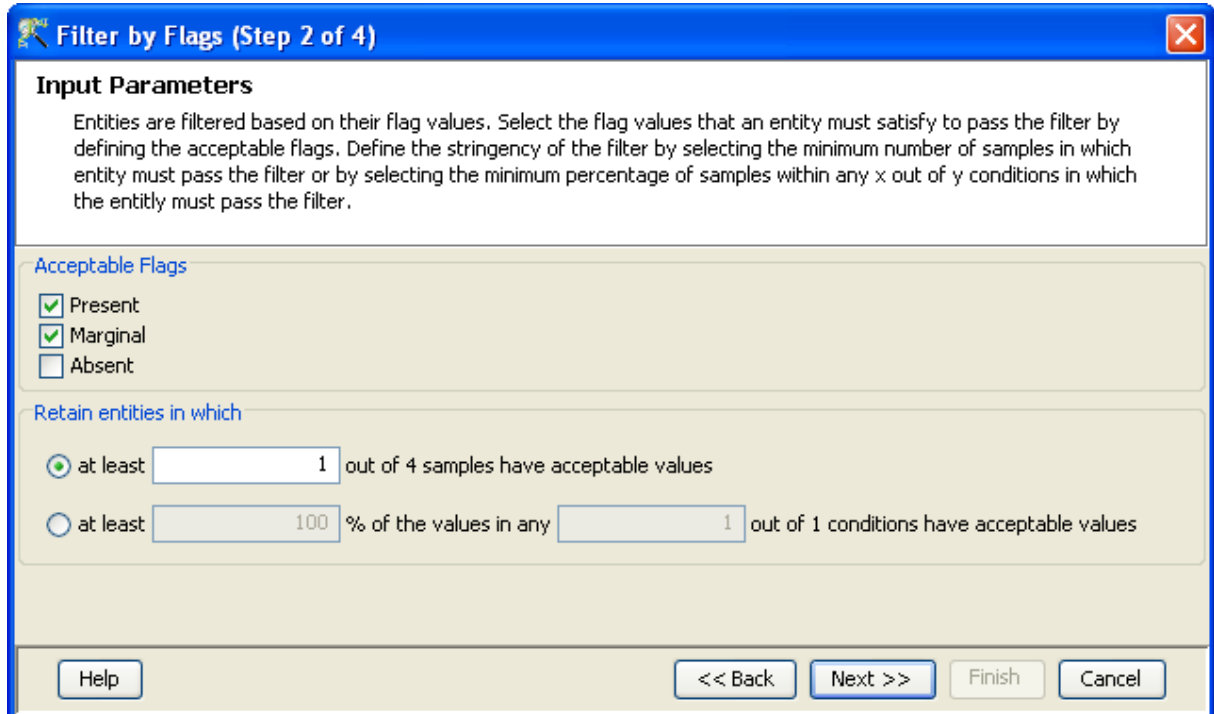


Figure 11.34: Input Parameters

Stats	FE Stats Used	Description/Measures
eQCOneColor LinFitLogLowConc	eQCOneColor LinFitLogLowConc	Log of lowest detectable concentration from fit of Signal vs. Concentration of E1a probes
AnyColorPrcent BGNonUnifOL	AnyColorPrcent BGNonUnifOL	Percentage of LocalBkgdRegions that are NonUnifOlr in either channel
gNonCtrlMedPrcent CVBGSub Sig	rNonCtrlMedPrcent CVBGSUB-Sig(red channel)	The median percent CV of background-subtracted signals for inlier noncontrol probes
gE1aMedCVBk SubSignal	geQCMedPrcentCVBG SubSig	Median CV of replicated E1a probes: Green Bkgd-subtracted signals
gSpatialDetrend RMSFilteredMinusFit	gSpatialDetrend RMSFilteredMinusFit	Residual of background detrending fit
absGE1E1aSlope	Abs(eQCOneColorLinFitSlope)	Absolute of slope of fit for Signal vs. Concentration of E1a probes
gNegCtrlAve BGSubSig	gNegCtrlAve BGSubSig	Avg of NegControl Bkgd-subtracted signals (Green)
gNegCtrlSDev BGSubSig	gNegCtrlSDev BGSubSig	StDev of NegControl Bkgd-subtracted signals (Green)
AnyColorPrcent FeatNonUnifOL	AnyColorPrcent FeatNonUnifOL	Percentage of Features that are NonUnifOlr

Table 11.10: Quality Controls Metrics

4. Step 4 of 4: Click *Next* to annotate and save the entity list. See Figure 11.36

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on [Filter Probesets on Data Files](#)
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

11.4.3 Analysis

- **Statistical Analysis**

For details refer to section [Statistical Analysis](#) in the advanced workflow.

- **Filter on Volcano Plot**

For details refer to section [Filter on Volcano Plot](#)

- **Fold Change**

For details refer to section [Fold Change](#)

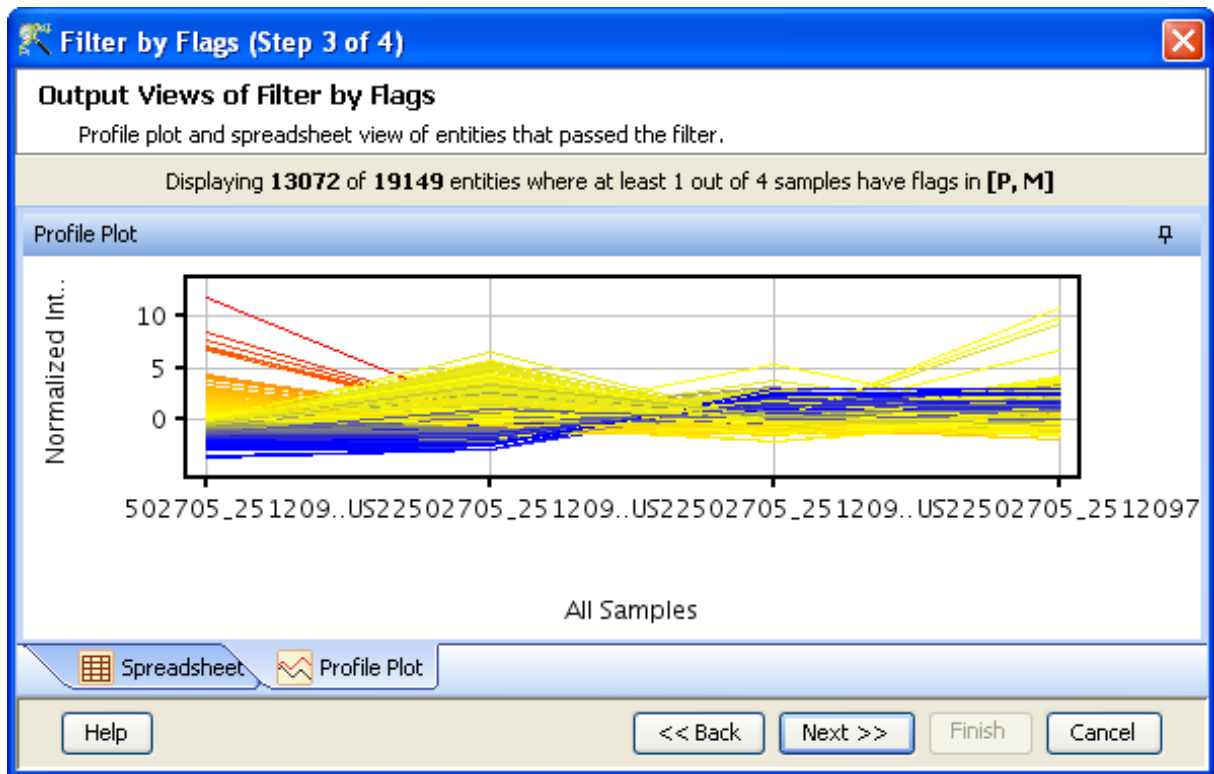


Figure 11.35: Output Views of Filter by Flags

- **Clustering**

For details refer to section [Clustering](#)

- **Find Similar Entities**

For details refer to section [Find Similar Entities](#)

- **Filter on Parameters**

For details refer to section [Filter on Parameters](#)

- **Principal Component Analysis**

For details refer to section [PCA](#)

11.4.4 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)

- **Run Prediction** For details refer to section [Run Prediction](#)

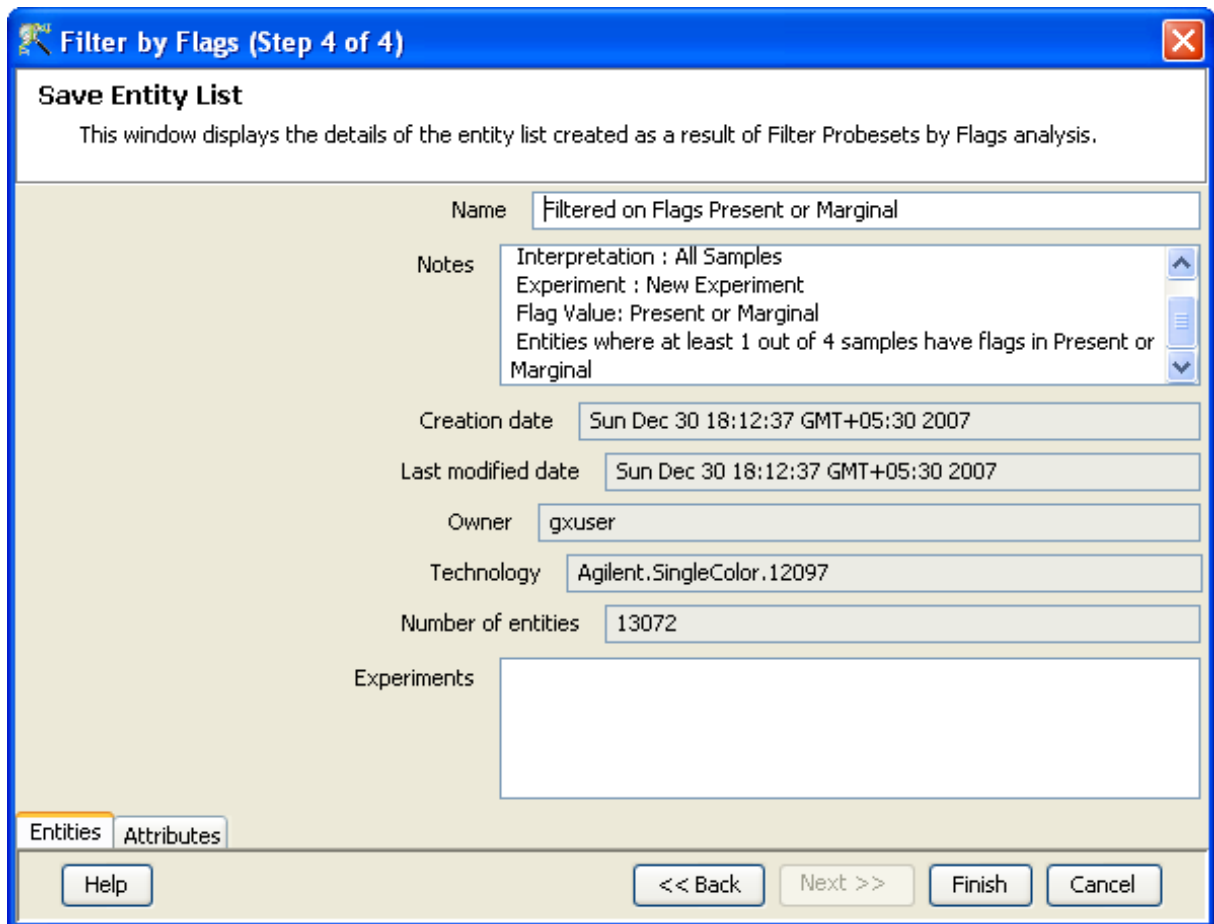


Figure 11.36: Save Entity List

11.4.5 Results

- **Gene Ontology (GO) analysis**

GO is discussed in a separate chapter called [Gene Ontology Analysis](#).

- **Gene Set Enrichment Analysis (GSEA)**

Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).

- **Gene Set Analysis (GSA)**

Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).

- **Pathway Analysis**

Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).

- **Find Similar Entity Lists**

This feature is discussed in a separate section called [Find Similar Entity Lists](#)

- **Find Significant Pathways**

This feature is discussed in a separate section called [Find Significant Pathways](#).

- **Launch IPA**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Import IPA Entity List**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Extract Interactions via NLP**

This feature is discussed in detail in the chapter [Pathway Analysis](#).

11.4.6 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)

- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)

- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)

- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

Chapter 12

Analyzing Agilent Two Color Expression Data

GeneSpring GX supports Agilent Two Color technology, with data files in .txt or .gpr formats. The data files in .txt format are obtained from Agilent Feature Extraction(FE) 8.5 and 9.5.3. When the data file is imported into **GeneSpring GX** the following columns get imported: ControlType, ProbeName, Signal (2 columns) and feature columns (2 sets). With files in .gpr formats, DesignID information is required; if present in the file, it is automatically recognized for import, or the user is prompted to input the DesignID. Note that if the design ID is not correct, there may be errors while processing the data.

Agilent Two Color Workflow supports most of the Standard Agilent technologies. The Agilent custom arrays other than .gpr formats, and the files from FE other than 8.5 and 9.5.3 can be analyzed by creating a Generic Two Color technology and using the corresponding workflow. In order to do so, certain column markings should be indicated (which are automatically done with standard technologies). These details can be found in the section on [Custom Agilent Arrays](#), while the Generic Two Color technology creation is available in Chapter 16 in the section [Creating Technology](#). Agilent Two Color files can be also split into single channels and analyzed as single color files. For the above situation, see the section on [Analyzing Agilent Two Color data in Agilent Single Color Experiment Type](#)

12.1 Running the Agilent Two Color Workflow

Upon launching **GeneSpring GX** , the startup is displayed with 3 options.

- Create new project
- Open existing project
- Open recent project



Figure 12.1: Welcome Screen

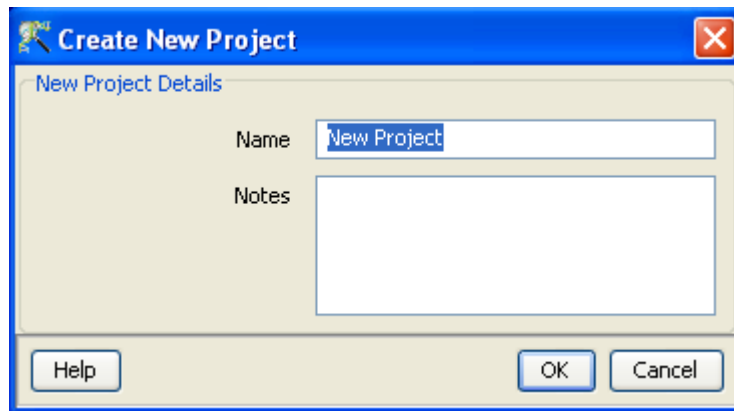


Figure 12.2: Create New project

Either a new project can be created or a previously generated project can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on **OK** to proceed.

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment:** This allows the user to create a new experiment. (steps described below).
2. **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

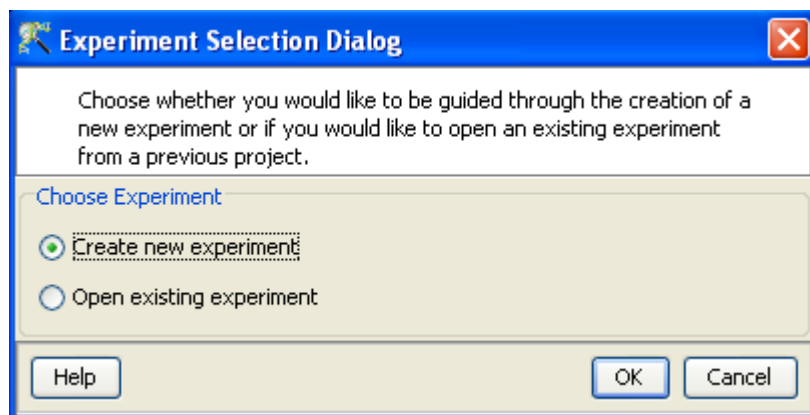


Figure 12.3: Experiment Selection

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the experiment type gives the user the option to choose between the multiple experiment types namely Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Agilent miRNA, Generic Single Color, Generic Two Color, Pathway and RealTime-PCR experiment.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Guided Workflow**
2. **Advanced Analysis Workflow**

Guided Workflow is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Advanced Analysis**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)
2. Choose Samples
3. Reorder
4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from

a scanner. On the other hand, a sample is created within **GeneSpring GX**, when it associates the data files with its appropriate technology (See the section on [Technology](#)). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the **Choose Samples** option. For selecting data files and creating an experiment, click on the **Choose File(s)** button, navigate to the appropriate folder and select the files of interest. Click on **OK** to proceed.

The technology specific for any chip type needs to be created or downloaded only once. Thus, upon creating an experiment of a specific chip type for the first time, **GeneSpring GX** prompts the user to download the technology from the update server. If the technology is not present, then **GeneSpring GX** creates it on the fly using user provided data identifiers. Annotations from a file can be added at any time by going to **Annotations**→**Update Technology Annotations**. If an experiment has been created previously with the same technology, **GeneSpring GX** then directly proceeds with experiment creation. Clicking on the **Choose Samples** button, opens a sample search wizard, with the following search conditions:

1. **Search field:** Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition:** Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on **Add** button, or can be removed from the list using **Remove** button.

Files can either be removed or reordered during the data loading step using the **Remove** or **Reorder** button.

Figures [12.4](#), [12.5](#), [12.6](#), [12.7](#) show the process of choosing experiment type, loading data, choosing samples and re-ordering the data files.

The next step gives the option of performing Dye-Swap on selected samples. Data/Sample files chosen in previous step are shown here and the user can select those arrays that were dye-swapped while performing the experiment. Accordingly, **GeneSpring GX** will swap the data between cy5 and cy3 for these arrays. (See Figure [12.8](#))

The *Guided Workflow* wizard appears with the sequence of steps on the left hand side with the current step being highlighted. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps.

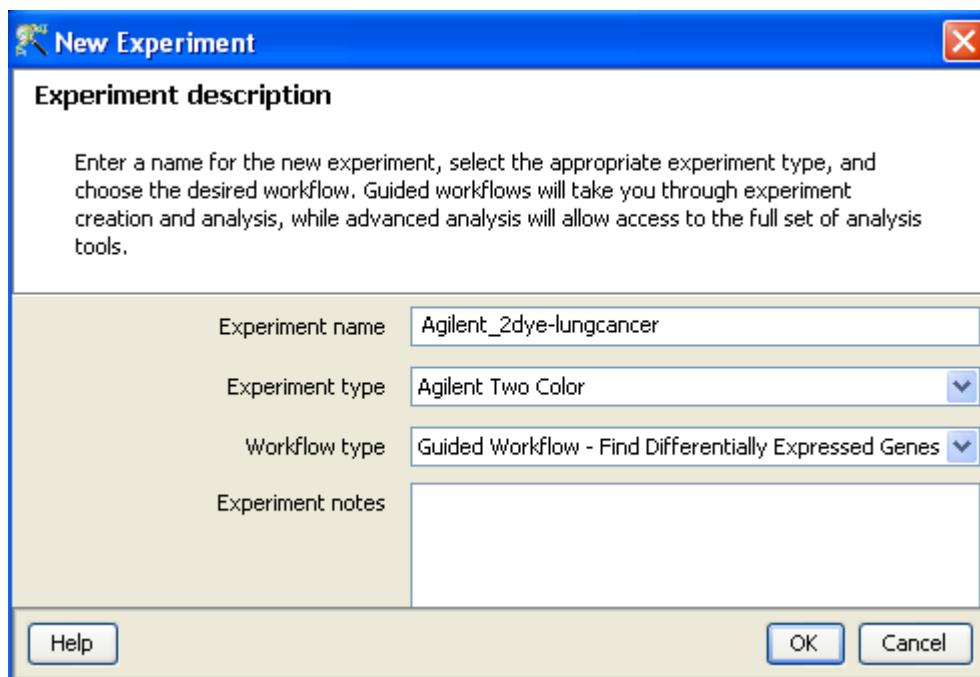


Figure 12.4: Experiment Description

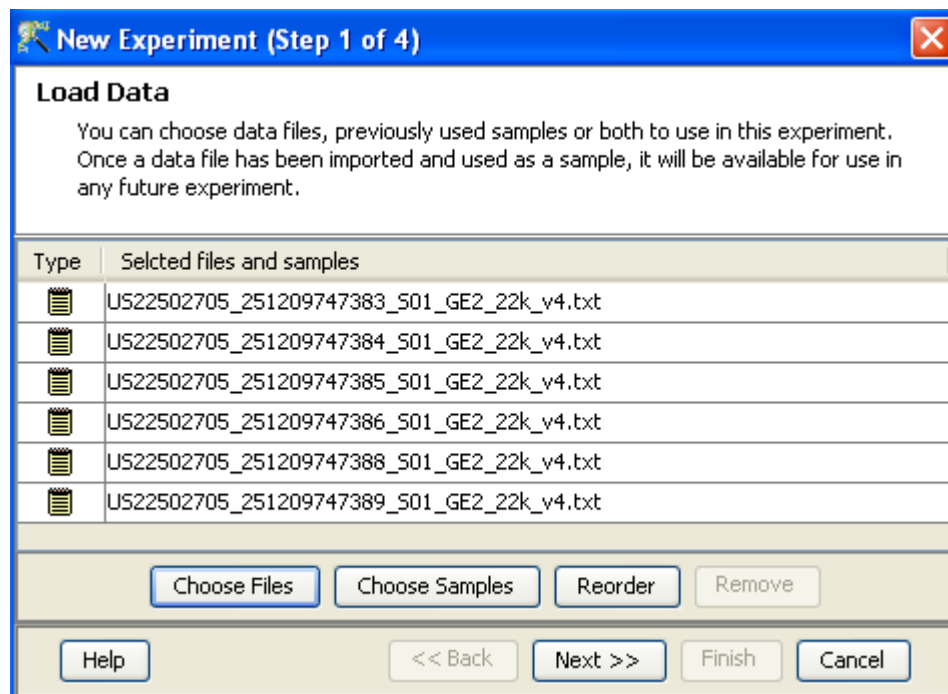


Figure 12.5: Load Data

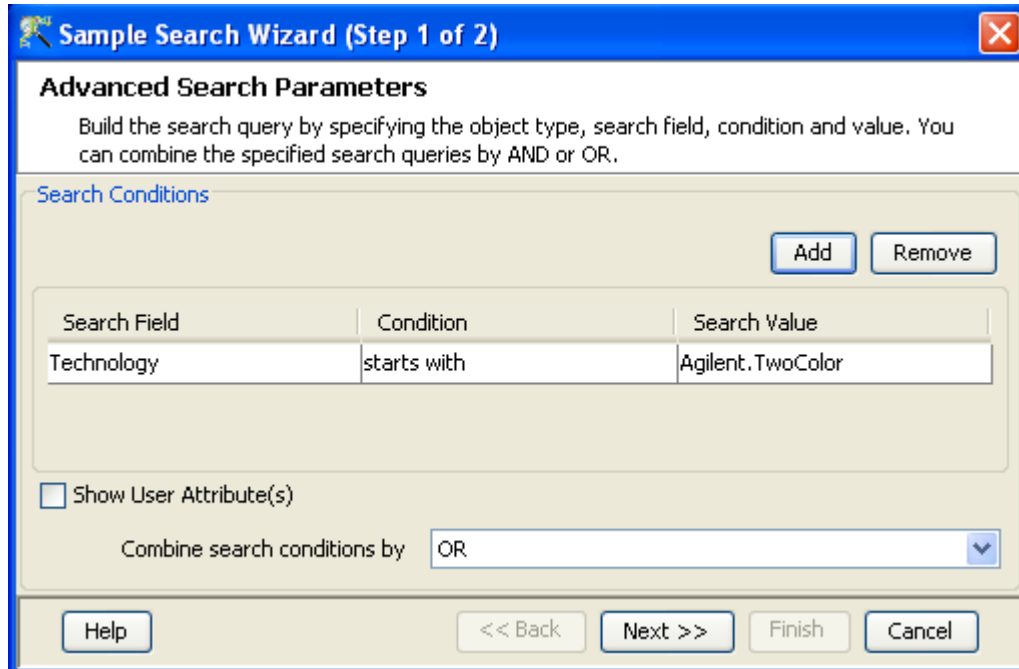


Figure 12.6: Choose Samples

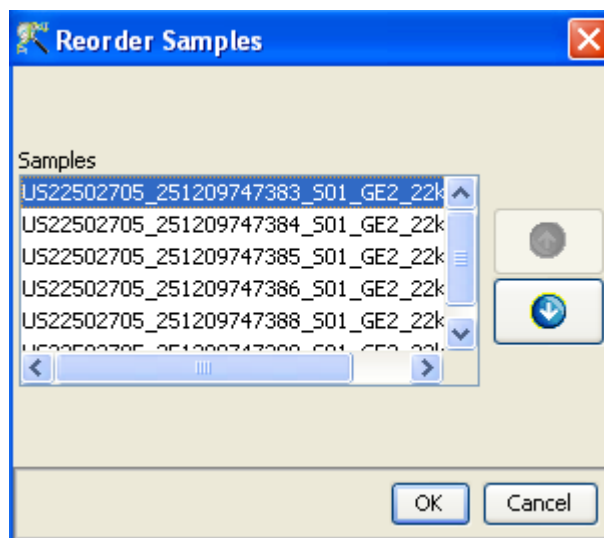


Figure 12.7: Reordering Samples

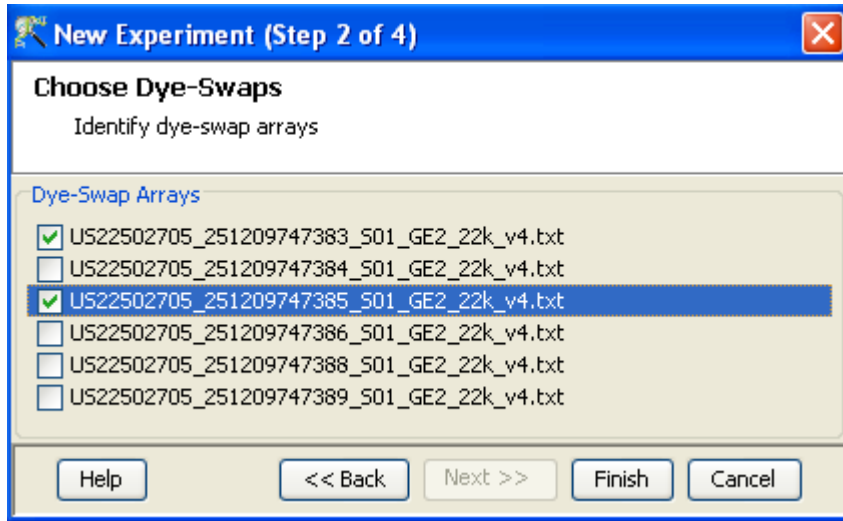


Figure 12.8: Dye Swap

12.2 Data Processing for Agilent Two Color arrays

- **File formats:** The data files should be in text (.txt) format (obtained from Agilent Feature Extraction (FE) 8.5 and 9.5.3) or in .gpr format.
- **Raw Signal Values:** The term "raw" signal values refer to the linear data after thresholding and summarization for the individual channels (cy3 and cy5). Summarization is performed by computing the geometric mean.
- **Normalized Signal Values:** The term Normalized signal value refers to the data after ratio computation, log transformation and Baseline Transformation.
- **Treatment of on-chip replicates:** For each replicate with multiple flags, the order of importance is Absent(A)>Marginal(M)>Present(P). If there is even one A, then the resultant flag is 'A'. If there is no A, but M and P, then M is assigned. If there are only Ps then only the resultant flag is assigned as 'P'. To get the overall flag for all replicates, **GeneSpring GX** excludes 'A' flag and assigns the majority considering the remaining ones. If there are only 'A' flags, only then the overall flag becomes 'A'. The following two examples illustrate this.
- **Flag values:** The flag value of a particular probeset is dependant on the flag values of the probes in it. If a probeset contains a probe which is marked as Present (P), the probeset is marked as P irrespective of the other flag values. The order of importance for flag values is Present>Marginal>Absent.
- **Treatment of Control probes:** The control probes are included while performing normalization. However there should be an exact match between the control probes in the technology and the sample for the probes to be utilized, as the comparison between the identifier columns is case-sensitive.
- **Empty Cells:** Not Applicable.
- **Sequence of events:** The sequence of events involved in the processing of the data files is: Thresholding→Dye swap→ratio computation→log transformation→Baseline Transformation.

	Signal	f1	f2	f3	(Resultant flag, A>M>P)
p1	1	P	M	A	A
p1	2	P	M	M	M
p1	3	P	P	P	P
p1	4	M	M	P	M
p1	5	M	P	M	M

Overall flag for p1 (Exclude A and assign majority) : M
Overall Signal = (2+3+4+5)/4 = 3.5

Figure 12.9: Agilent Two Colour - Handling on chip replicates: Example 1

	Signal	f1	f2	f3	(Resultant flag, A>M>P)
p1	1	P	M	A	A
p1	2	A	M	P	A
p1	3	M	A	P	A
p1	4	A	A	P	A
p1	5	A	A	A	A

Overall flag for p1 (No P or M present, so take A) : A
Overall Signal = (1+2+3+4+5)/5 = 3

Figure 12.10: Agilent Two Colour - Handling on chip replicates: Example 2

12.3 Guided Workflow steps

Summary report (Step 1 of 8): The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized Expression values on the Y axis. An information message on the top of the wizard shows the number of samples in the file and the sample processing details. If the number of samples are more than 30, they are only represented in a tabular column. On clicking the *Next* button it will proceed to the next step and on clicking *Finish*, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. On doing a right click, the options of invert selection is displayed and on clicking the same the selection is inverted i.e., all the probes except the selected ones are highlighted in green. Figure 12.11 shows the Summary report with box-whisker plot.

Note: In the *Guided Workflow*, these default parameters cannot be changed. To choose different parameters use *Advanced Analysis*.

Experiment Grouping (Step 2 of 8): On clicking *Next*, the *Experiment Grouping* window appears which is the 2nd step in the **Guided Workflow**. It requires parameter values to be defined to

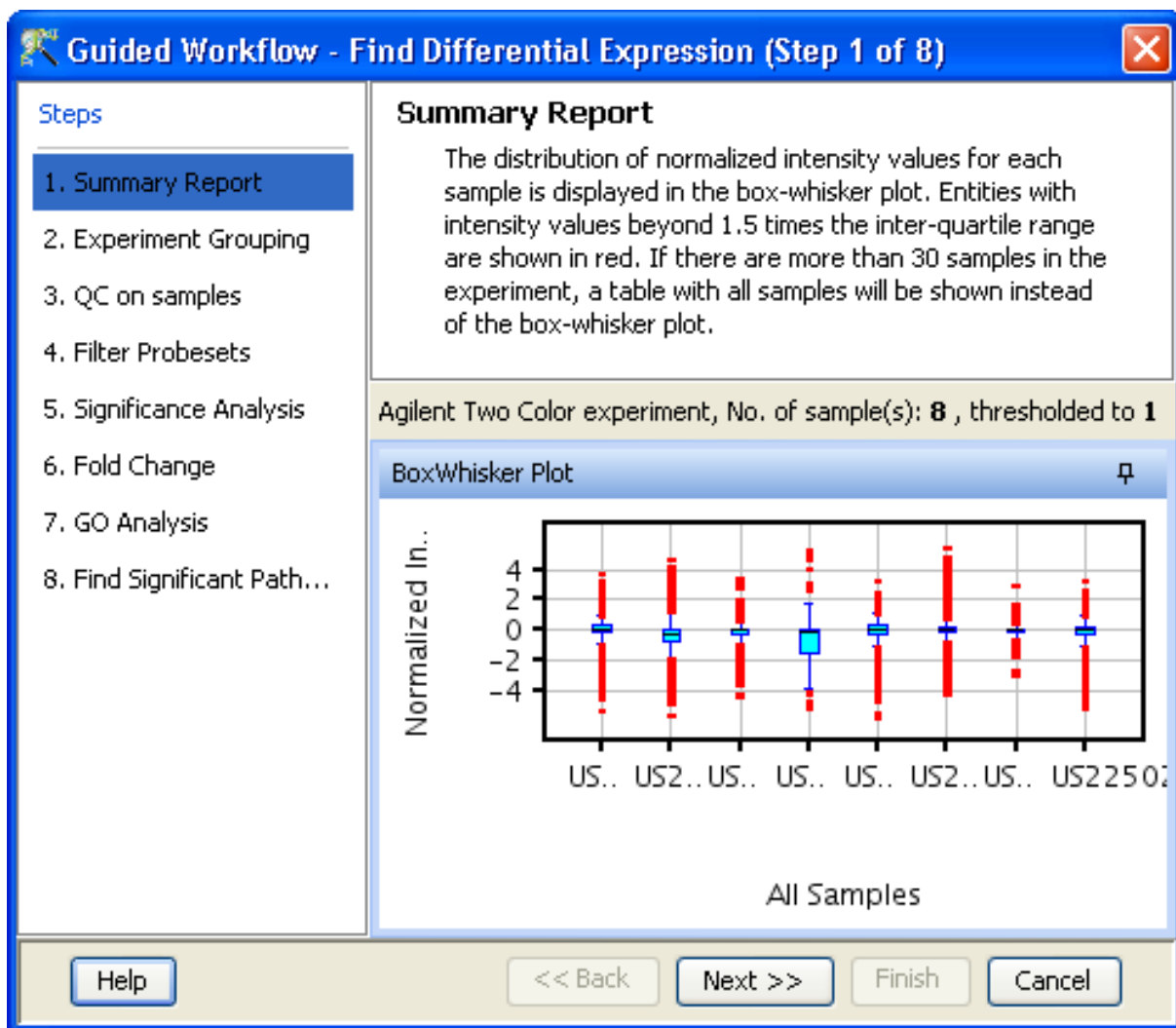




Figure 12.11: Summary Report

group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the **Add parameter** button. Parameter values can be assigned by first selecting the desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on **Clear**. Press **OK** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Guided Workflow**. The other parameters can be used in the **Advanced Analysis**.

Note: The *Guided Workflow* does not proceed further without grouping information.





Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file  icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file

should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

```
A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50
```

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file  icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click → *Properties* → *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click → *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

Note: The *Guided Workflow* by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the guided wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures [12.12](#) and [12.13](#) respectively.

Quality Control (Step 3 of 8): The 3rd step in the Guided workflow is the QC on samples which is displayed in the form of four tiled windows.

Note that for experiments created using .gpr file formats, the *Quality Control* step is skipped.

The four tiled windows are as follows:

- Quality controls Metrics- Report and Experiment grouping tabs
- Quality controls Metrics- Plot
- PCA scores
- Legend

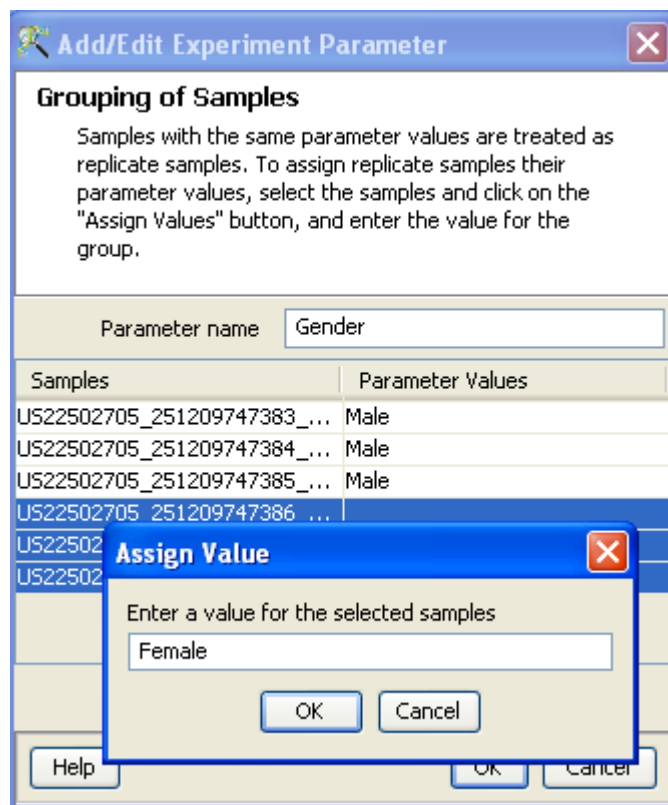


Figure 12.12: Experiment Grouping

QC on Samples generates four tiled windows as seen in Figure 12.14.

The metrics report include statistical results to help you evaluate the reproducibility and reliability of your microarray data.

The table shows the following:

More details on this can be obtained from the Agilent Feature Extraction Software(v9.5) Reference Guide, available from <http://chem.agilent.com>.

Quality controls Metrics Plot shows the QC metrics present in the QC report in the form of a plot. *Principal Component Analysis (PCA)* calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed on the samples. Click on *OK* to proceed.

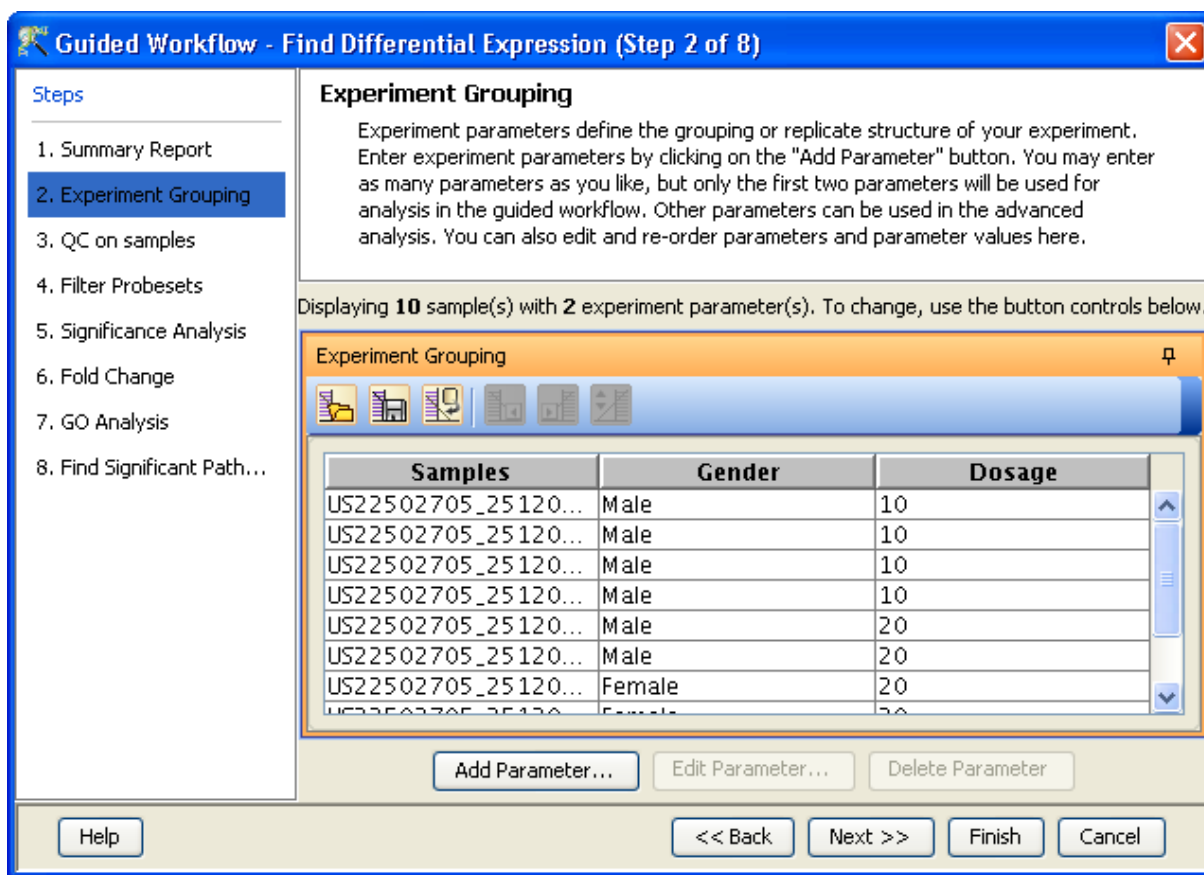


Figure 12.13: Edit or Delete of Parameters

The fourth window shows the legend of the active QC tab.

Filter probesets (Step 4 of 8): In this step, the entities are filtered based on their flag values P (*present*), M (*marginal*) and A (*absent*). Only entities having the present and marginal flags in at least one sample are displayed as a profile plot. The selection can be changed using Rerun Filter option. The flagging information is derived from the Feature columns in data file. More details on how flag values [P,M,A] are calculated can be obtained from *QC Chart Tool* and <http://www.chem.agilent.com>. The plot is generated using the normalized signal values and samples grouped by the active interpretation. Options to customize the plot can be accessed via the Right-click menu. An *Entity List*, corresponding to this filtered list, will be generated and saved in the Navigator window. The Navigator window can be viewed after exiting from *Guided Workflow*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Newer annotations can be added and existing ones removed using the *Configure Columns* button. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The cutoff for filtering can be changed using the *Rerun Filter* button. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. Double click on *Profile Plot* opens up an entity inspector giving the annotations corresponding to the selected profile. The information message on the top shows the number of entities satisfying the flag values. Figures 12.15 and 12.16 are displaying the profile plot obtained in situations having single and two parameters.

Significance Analysis (Step 5 of 8) Depending upon the experimental grouping, **GeneSpring GX**

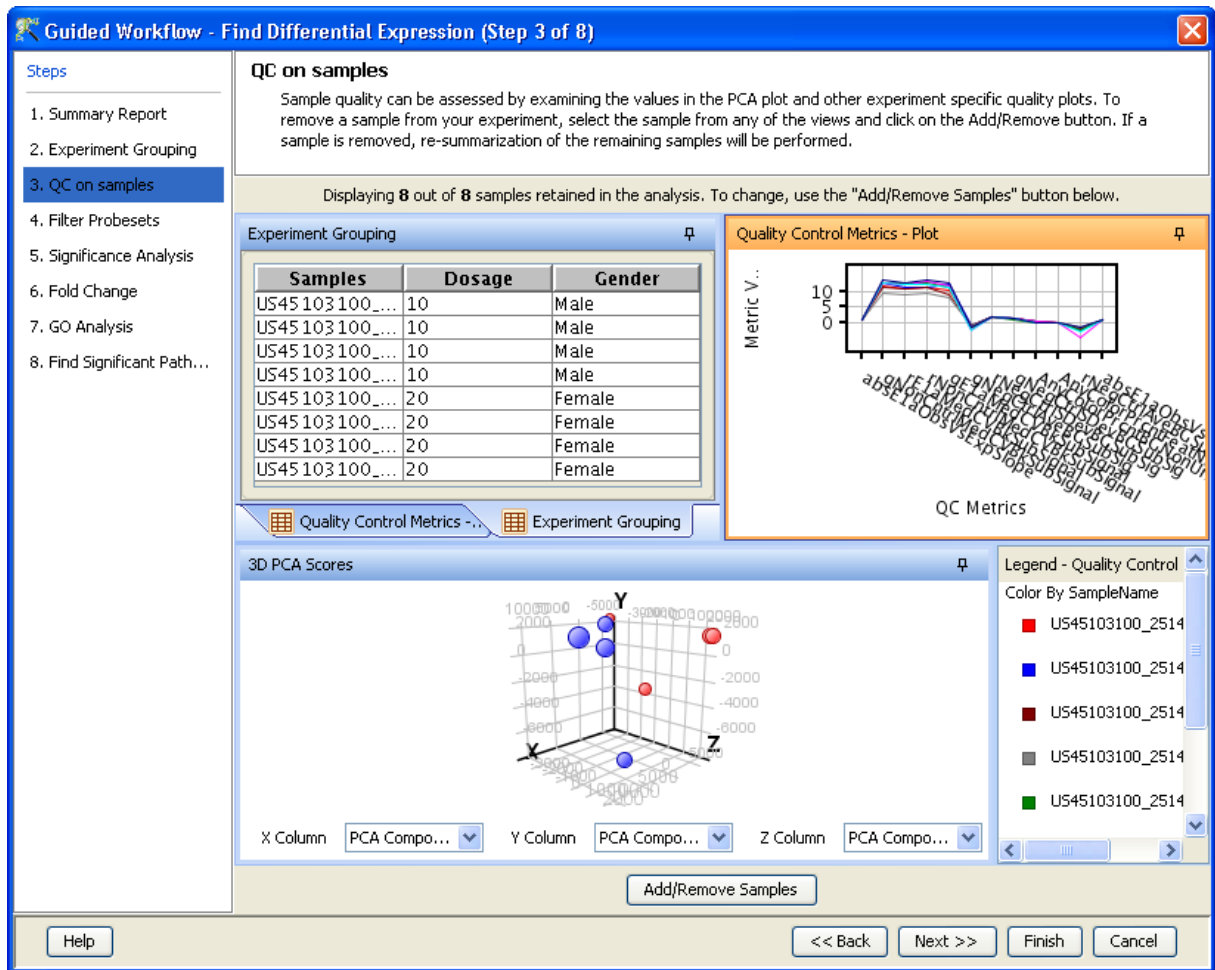


Figure 12.14: Quality Control on Samples

performs either T-test or ANOVA. The tables below describe broadly the type of statistical test performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I*, has 2 groups, the normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.
- **Example Sample Grouping II:** In this example, only one group, the tumor, is present. T-test against zero will be performed here.
- **Example Sample Grouping III:** When 3 groups are present (normal, tumor1 and tumor2) and one of the groups (tumor2 in this case) does not have replicates, statistical analysis cannot be performed. However if the condition tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.
- **Example Sample Grouping IV:** When there are 3 groups within an interpretation, One-way ANOVA will be performed.
- **Example Sample Grouping V:** This table shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and

Name of Metric	FE Stats Used	Description/Measures
absE1aObsVs ExpSlope	Abs(eQCObsVs ExpLRSlope)	Absolute of slope of fit for Observed vs. Expected E1a LogRatios
gNonCntrlMedCVBk SubSignal	gNonCntrlMedCVBk SubSignal	Median CV of replicated Non-Control probes: Green Bkgd-subtracted signals
rE1aMedCVBk SubSignal	reQCMedPrnt CVBGSubSig	Median CV of replicated E1a probes: Red Bkgd-subtracted signals
rNonCntrlMedCVBk SubSignal	rNonCntrlMedCVBk SubSignal	Median CV of replicated NonControl probes: Red Bkgd-subtracted signals
gE1aMedCVBk SubSignal	geQCMedPrnt CVBGSubSig	Median CV of replicated E1a probes: Green Bkgd-subtracted signals
gNegCtrlAve BGSubSig	gNegCtrlAve BGSubSig	Avg of NegControl Bkgd-subtracted signals (Green)
rNegCtrlAve BGSubSig	rNegCtrlAve BGSubSig	Avg of NegControl Bkgd-subtracted signals (Red)
gNegCtrlSDev BGSubSig	gNegCtrlSDev BGSubSig	StDev of NegControl Bkgd-subtracted signals (Green)
rNegCtrlSDevBGSUBSIG	rNegCtrlSDevBGSUBSIG	StDev of NegControl Bkgd-subtracted signals (Red)
AnyColorPrnt BGNonUnifOL	AnyColorPrnt BGNonUnifOL	Percentage of LocalBkgdRegions that are NonUnifOlr in either channel
AnyColorPrnt FeatNonUnifOL	AnyColorPrnt FeatNonUnifOL	Percentage of Features that are NonUnifOlr in either channel
absE1aObsVs ExpCorr	Abs(eQCObsVs ExpCorr)	Absolute of correlation of fit for Observed vs. Expected E1a LogRatios

Table 12.1: Quality Controls Metrics

Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.

- **Example Sample Grouping VI:** In this table, a two-way ANOVA will be performed.
- **Example Sample Grouping VII:** In the example below, a two-way ANOVA will be performed and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Statistical Tests: T-test and ANOVA

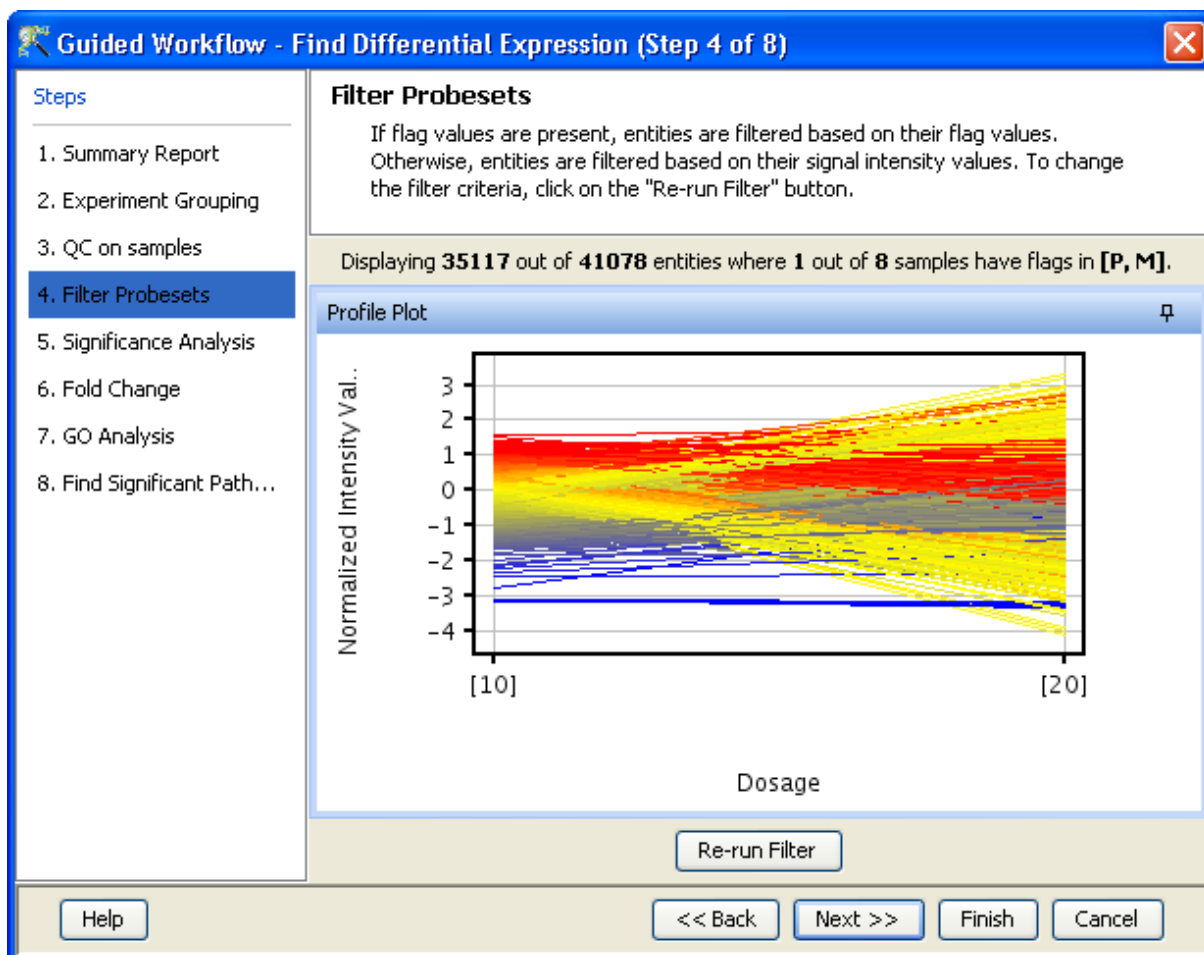


Figure 12.15: Filter Probesets-Single Parameter

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 12.2: Sample Grouping and Significance Tests I

- **T-test: T-test unpaired** is chosen as a test of choice with a kind of experimental grouping shown in Table 1. Upon completion of T-test the results are displayed as three tiled windows.
 - A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *Regulation*.
 - *Differential expression analysis report* mentioning the Test description i.e. test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic* or *Permutative*).

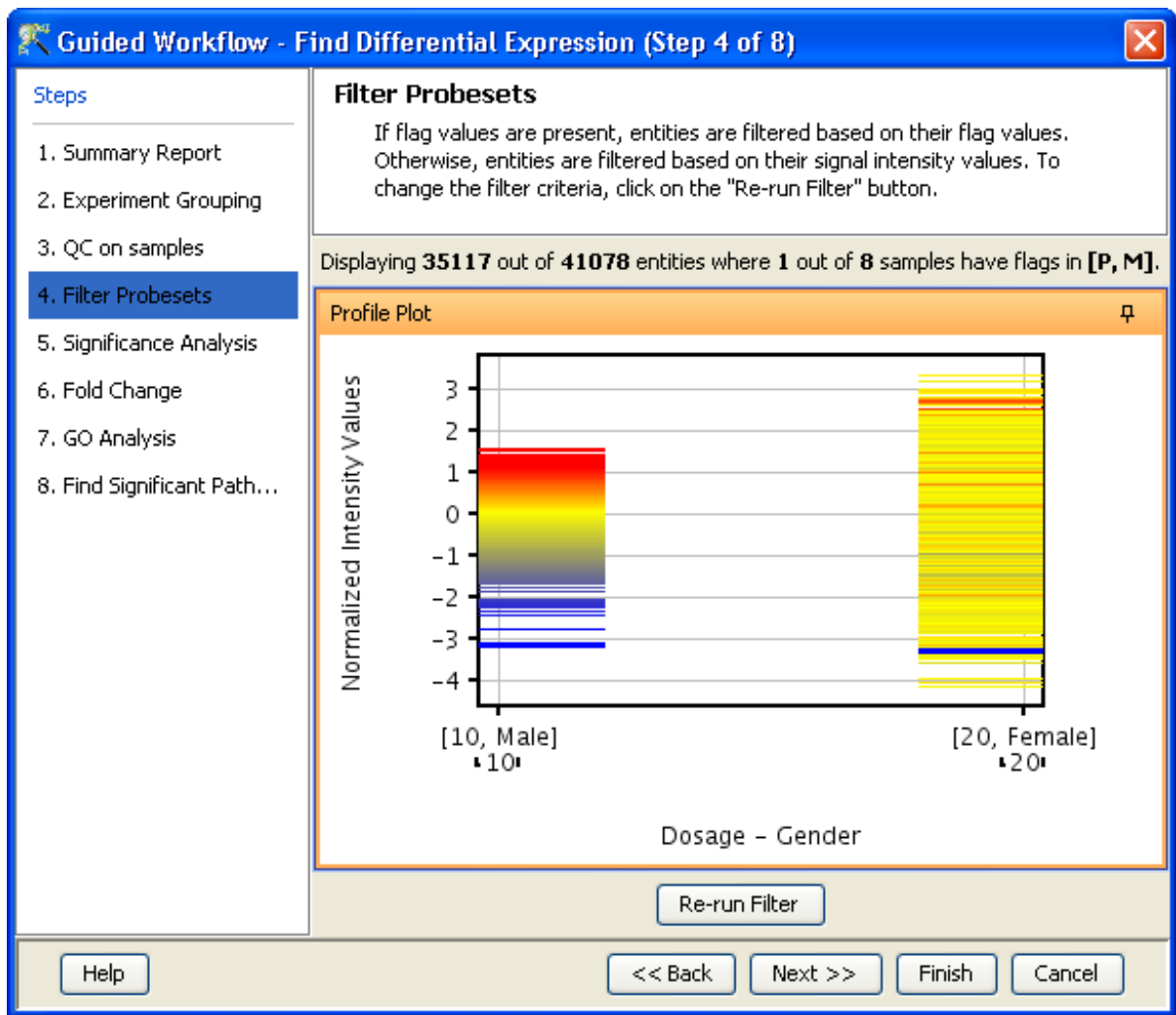


Figure 12.16: Filter Probesets-Two Parameters



Figure 12.17: Rerun Filter

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 12.3: Sample Grouping and Significance Tests II

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 12.4: Sample Grouping and Significance Tests III

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

- **Analysis of variance(ANOVA):** ANOVA is chosen as a test of choice under the experimental grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:
- A *p-value table* consisting of probe names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.
- *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and p-value computation type (*Asymptotic or Permutative*).
- *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

Special case: In situations when samples are not associated with at least one possible permutation of conditions (like Normal at 50 min and Tumor at 10 min mentioned above), no p-value can be computed and the **Guided Workflow** directly proceeds to **GO analysis**.

Fold-change (Step 6 of 8): Fold change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 12.5: Sample Grouping and Significance Tests IV

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 12.6: Sample Grouping and Significance Tests V

gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis. The wizard shows a table consisting of 3 columns: Probe Names, Fold change value and regulation (up or down). The regulation column depicts which one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using *Re-run Filter*. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1. A profile plot is also generated. Upregulated entities are shown in red. The color can be changed using the Right-click→*Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cutoff in the experiment Navigator.

Note: Fold Change step is skipped and the *Guided Workflow* proceeds to the *GO Analysis* in case of experiments having 2 parameters.

Fold Change view with the spreadsheet and the profile plot is shown in Figure 12.20.

Gene Ontology Analysis (Step 7 of 8): The *GO Consortium* maintains a database of controlled vocabularies for the description of molecular function, biological process and cellular location of gene products. The GO terms are displayed in the Gene Ontology column with associated *Gene Ontology Accession* numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 12.7: Sample Grouping and Significance Tests VI

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 12.8: Sample Grouping and Significance Tests VII

The GO analysis wizard shows two tabs comprising of a spreadsheet and a *GO tree*. The *GO Spreadsheet* shows the *GO Accession* and *GO terms* of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However, the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.1 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button is also provided on the GO tree panel to search using some keywords

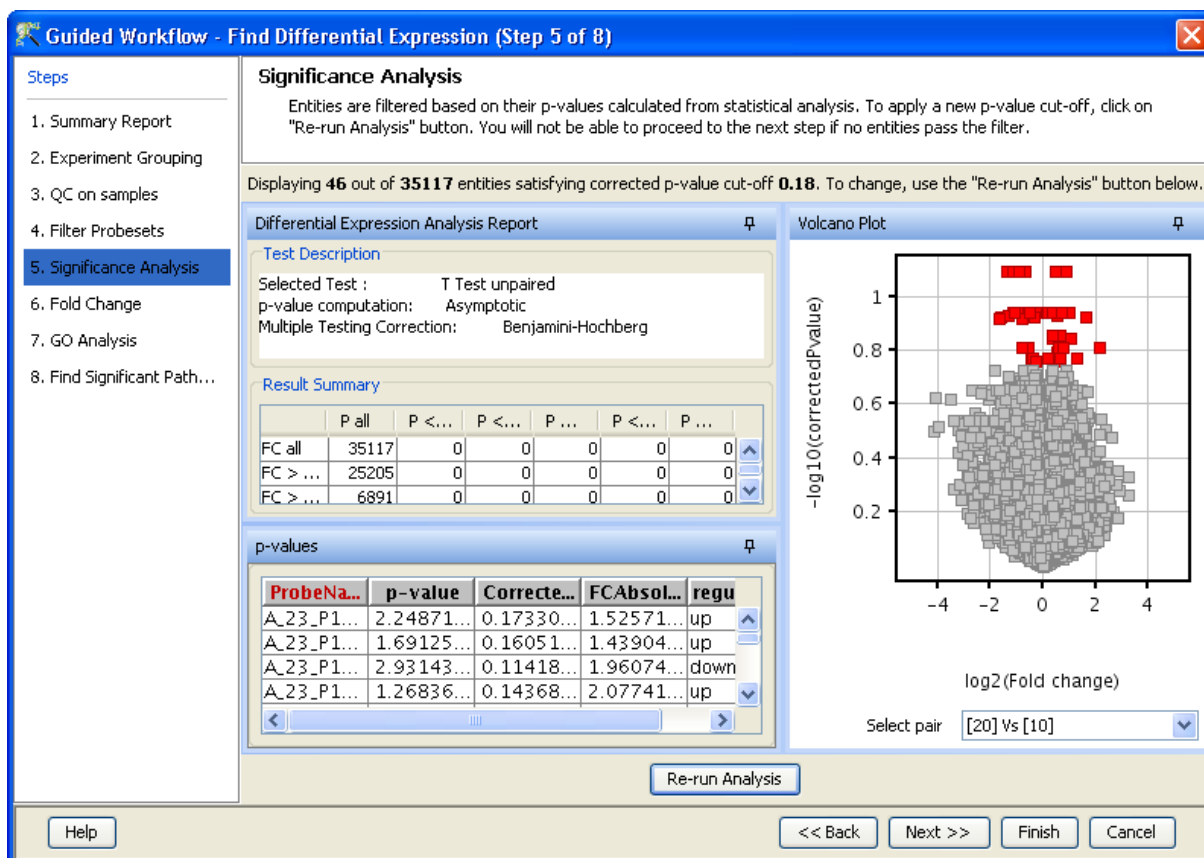


Figure 12.18: Significance Analysis-T Test

Note : In **GeneSpring GX** GO analysis implementation, all the three component: Molecular Function, Biological Processes and Cellular location are considered together.

On finishing the GO analysis, the *Advanced Workflow* view appears and further analysis can be carried out by the user. At any step in the Guided workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any) and the *Advanced Workflow* view appears.

Find Significant Pathways (Step 8 of 8): This step in the Guided Workflow finds relevant pathways from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the Fold Change (step 6 of 8). This view shows two tables-

- The Significant Pathways table shows the names of the pathways as well as the number of nodes and entities in the pathway and the p-values. It also shows the number of entities that are similar to the pathway and the entity list. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.
- The Non-significant Pathways table shows the pathways in the tool that do not have a single entity in common with the ones in the given entity list.

The user has an option of changing the p-value cut-off(using *Change cutoff*) and also to save specific pathways using the *Custom Save* option. See Figure 12.22. On clicking, *Finish* the main

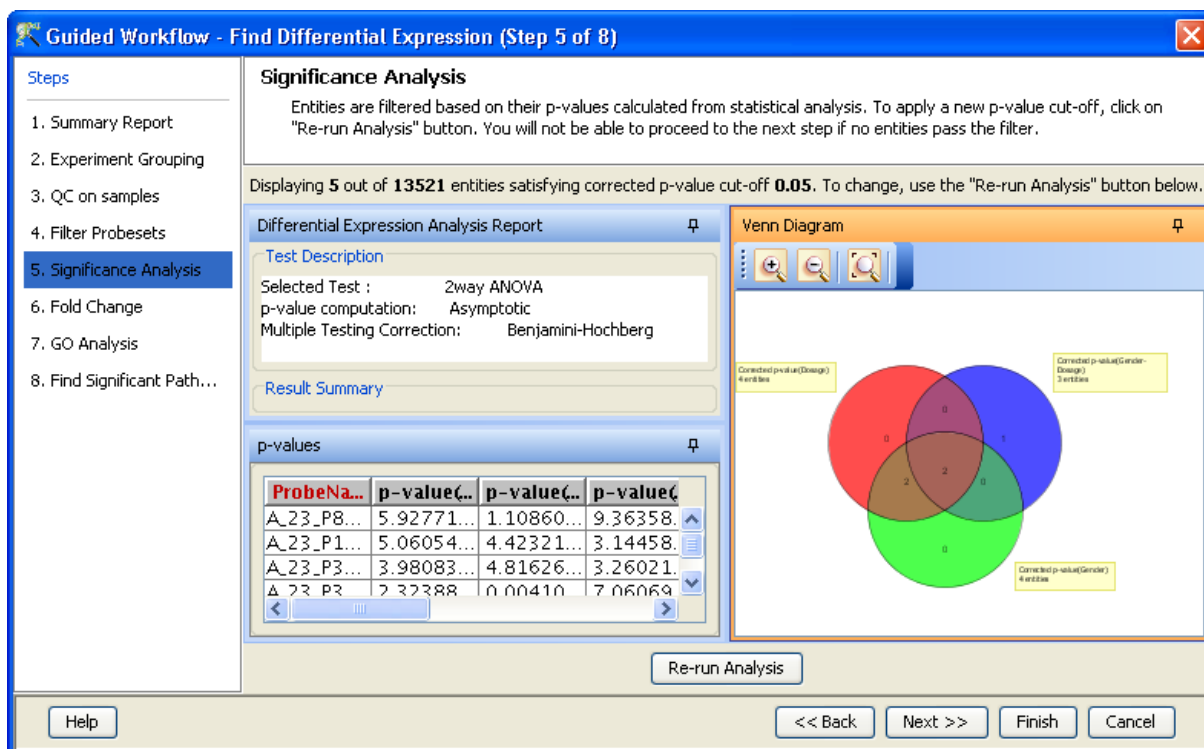


Figure 12.19: Significance Analysis-Anova

tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of the Guided Workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in the Guided Workflow, on clicking **Finish**, the analysis stops at that step (creating an entity list if any).

Note: In case the user is using **GeneSpring GX** for the first time, this option will give results using the demo pathways. The user can upload the pathways of his/her choice by using the option **Import BioPax pathways** under **Tools** in the **Menu** bar. Later instead of reverting to the Guided Workflow the user can use the option **Find Significant Pathways** in **Results Interpretation** under the same Workflow.

The default parameters used in the Guided Workflow is summarized below

12.4 Advanced Workflow

The *Advanced Workflow* offers a variety of choices to the user for the analysis. Flag options can be changed and raw signal thresholding can be altered. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Advanced Workflow*, load the data as described earlier. In the *New Experiment Dialog*, choose the Workflow Type as *Advanced Analysis*. Click *OK* will open a new experiment wizard which then proceeds as follows:

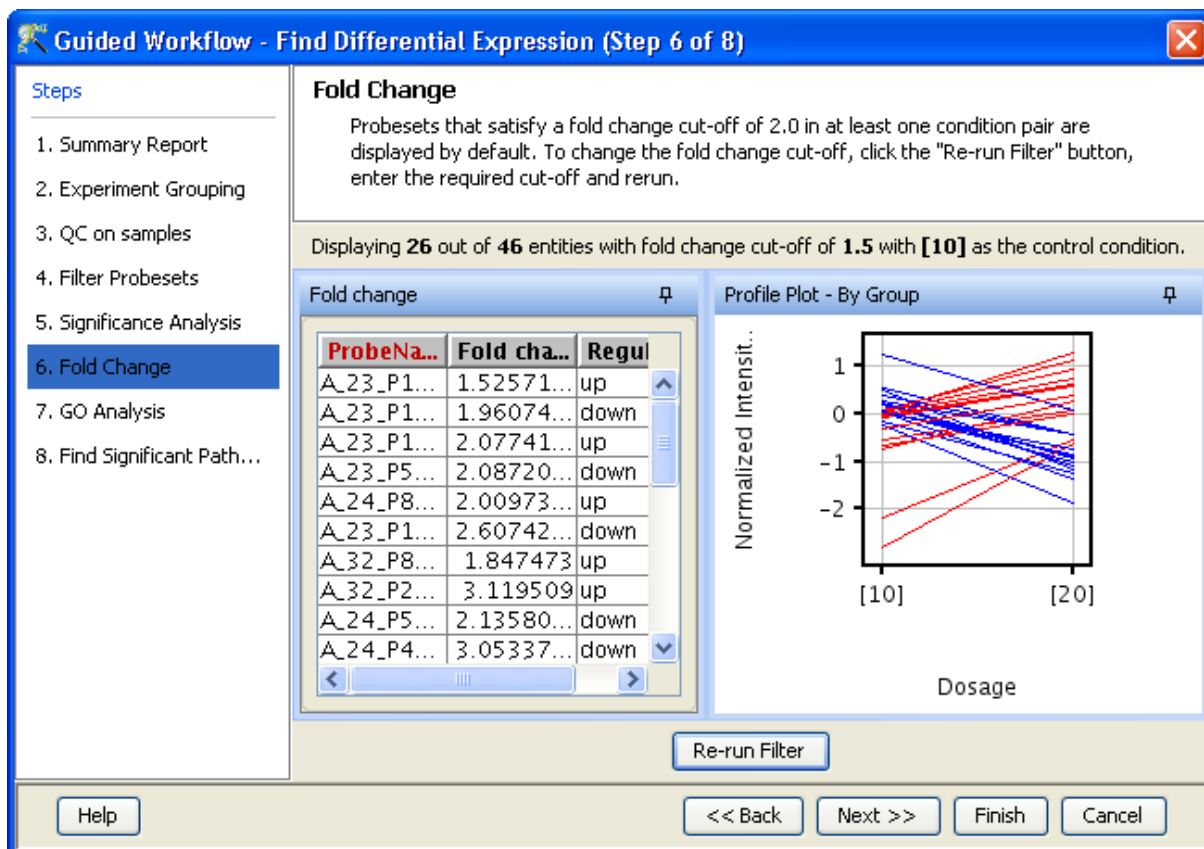


Figure 12.20: Fold Change

1. **Step 1 of 5: Load Data** As in case of *Guided Workflow*, either data files can be imported or else pre-created samples can be used.
 - For loading new txt files or gpr files, use Choose Files.
 - If these data files have been previously used in **GeneSpring GX** experiments *Choose Samples* can be used.

The *Load Data* window is shown in Figure 12.23.

2. **Step 2 of 5 : Samples Validation**

This step is shown only if there is mismatch in technology between the gpr files input in step 1. **GeneSpring GX** requires that the files input for any particular experiment be of the same technology. The work around is go back to step 1 and remove those sample files that are of different technology.

The *Samples Validation* window is shown in Figure 12.24.

3. **Step 3 of 5: Choose Dye-swaps** Dye-Swap arrays, if any, can be identified, in this step.

The *Choose Dye Swaps* window is depicted in the Figure 12.25.

4. **Step 4 of 5: Advanced Flag Import**

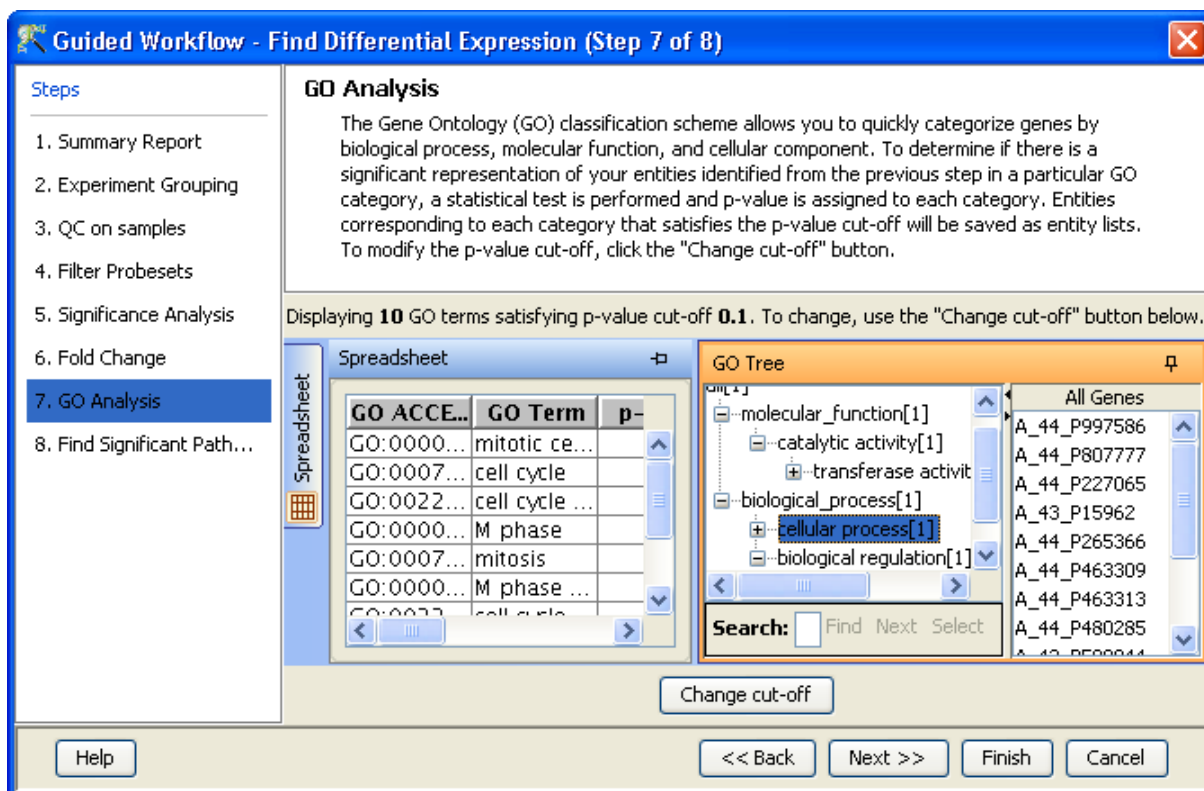


Figure 12.21: GO Analysis

This gives the options for importing flag information. The information is derived from the Feature columns in data file. User has the option of changing the default flag settings that appear in this step. The 'Save as Default' handle allows saving the current flag settings under the tool configuration. When a file is imported, **GeneSpring GX** will show these saved default setting in this step, by default. The settings can be changed either in this wizard or from *Tools* → *Options* → *Miscellaneous* → *Agilent Flag Settings*.

This step is skipped for files in .gpr formats.

Figure 12.26 shows the Step to import flags in Experiment Creation.

5. Step 5 of 5 : Preprocess Baseline Options

The final step of Experiment Creation is shown in Figure 6.24.

Criteria for preprocessing of input data is set here. It allows the user to choose the appropriate baseline transformation option.

The baseline options include:

- *Do not perform baseline*

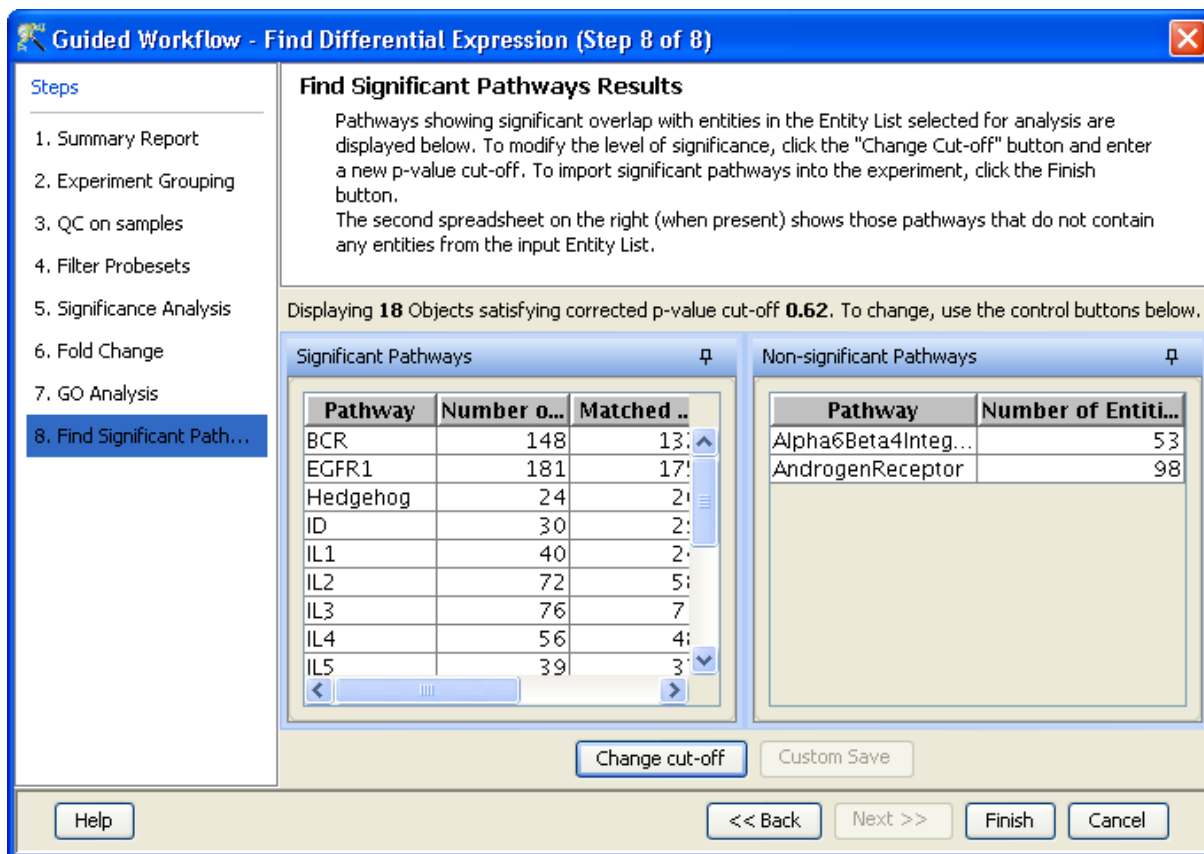


Figure 12.22: Find Significant Pathways

- **Baseline to median of all samples:** For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- **Baseline to median of control samples:** For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on **Assign value**. This opens up the **Choose Control Samples** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on **Ok**. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

12.4.1 Experiment Setup

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the Advanced Workflow,

	Parameters	Parameter values
Expression Data Transformation	Thresholding	1.0
	Normalization	Not Applicable
	Baseline Transformation	Not Applicable
	Summarization	Not Applicable
Filter by		
1.Flags	Flags Retained	Present(P), Marginal(M)
2.Expression Values	(i) Upper Percentile cutoff	Not Applicable
	(ii) Lower Percentile cutoff	
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1
Find Significant Pathways	p-value cutoff	0.05

Table 12.9: Table of Default parameters for Guided Workflow

the method of analysis, the details of the algorithms used and the interpretation of results

- **Experiment Grouping:** Experiment parameters defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index

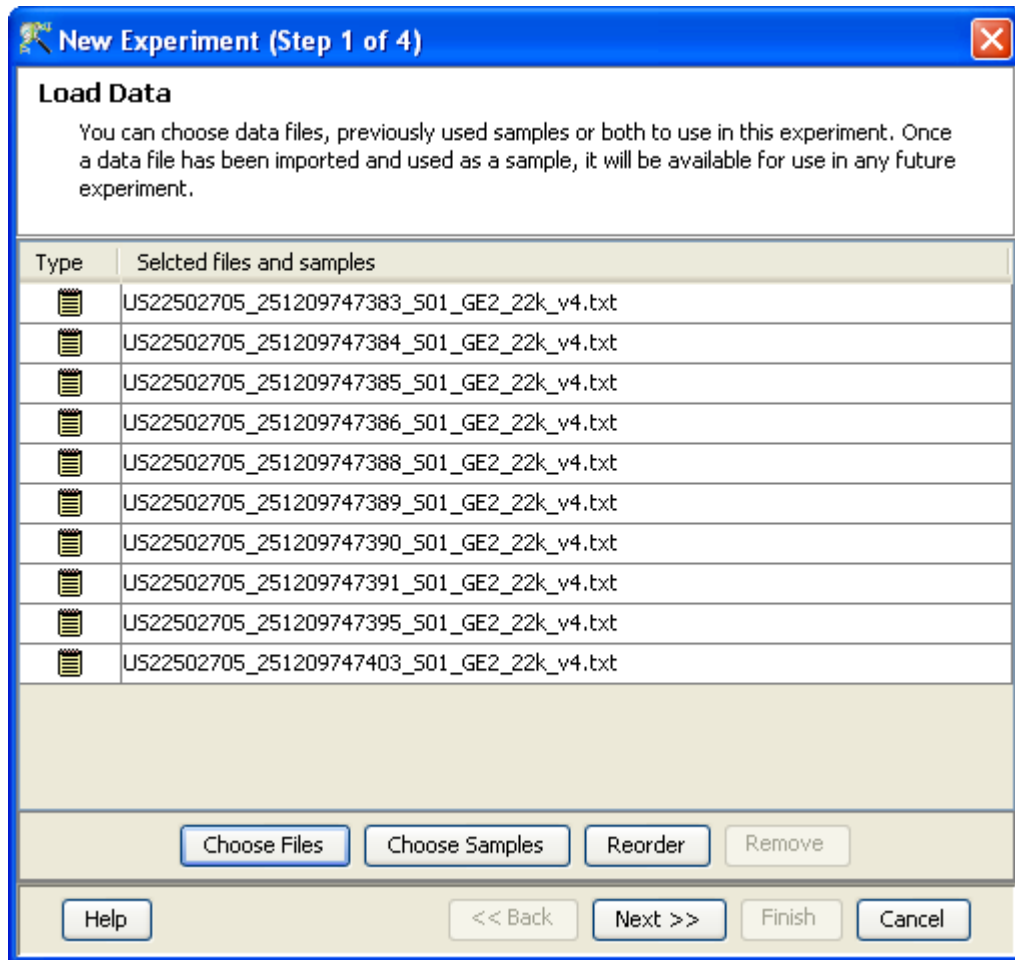


Figure 12.23: Load Data

- * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
 - Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
 - Experiment grouping, if present in the original experiment, will be retained.
 - The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed

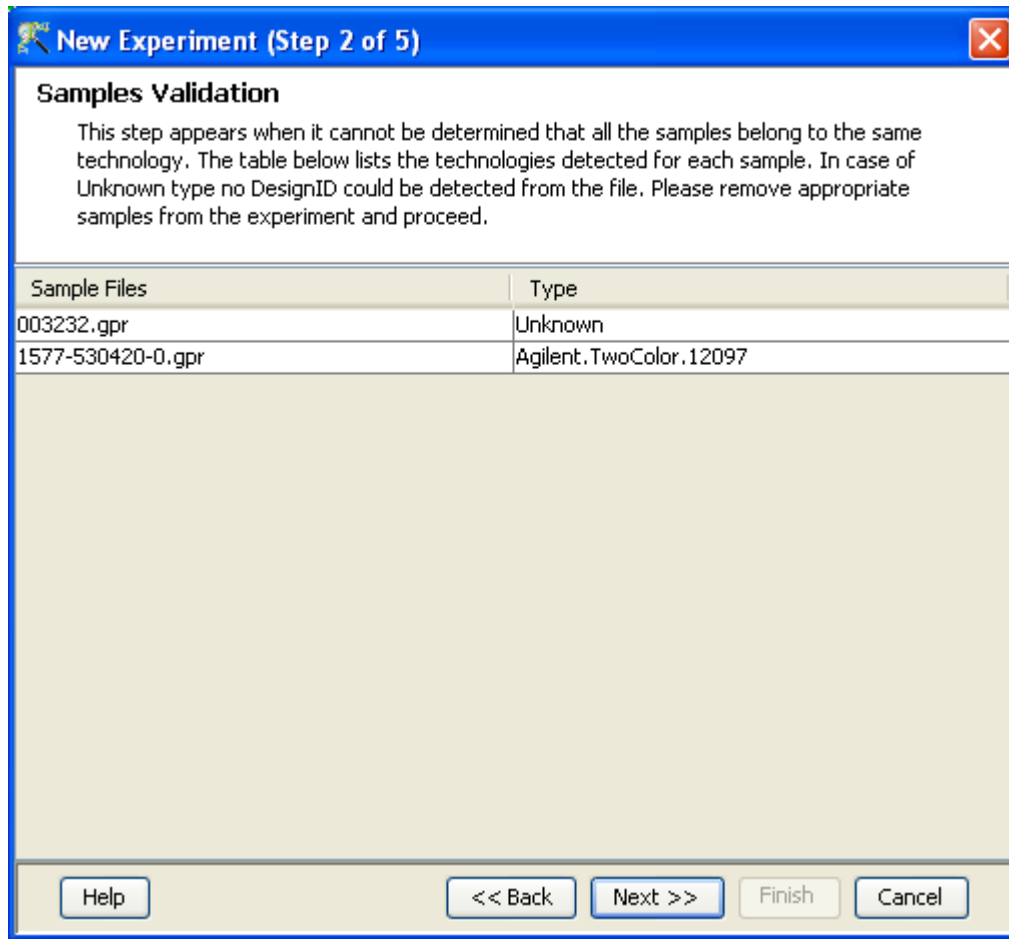


Figure 12.24: Samples Validation

out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

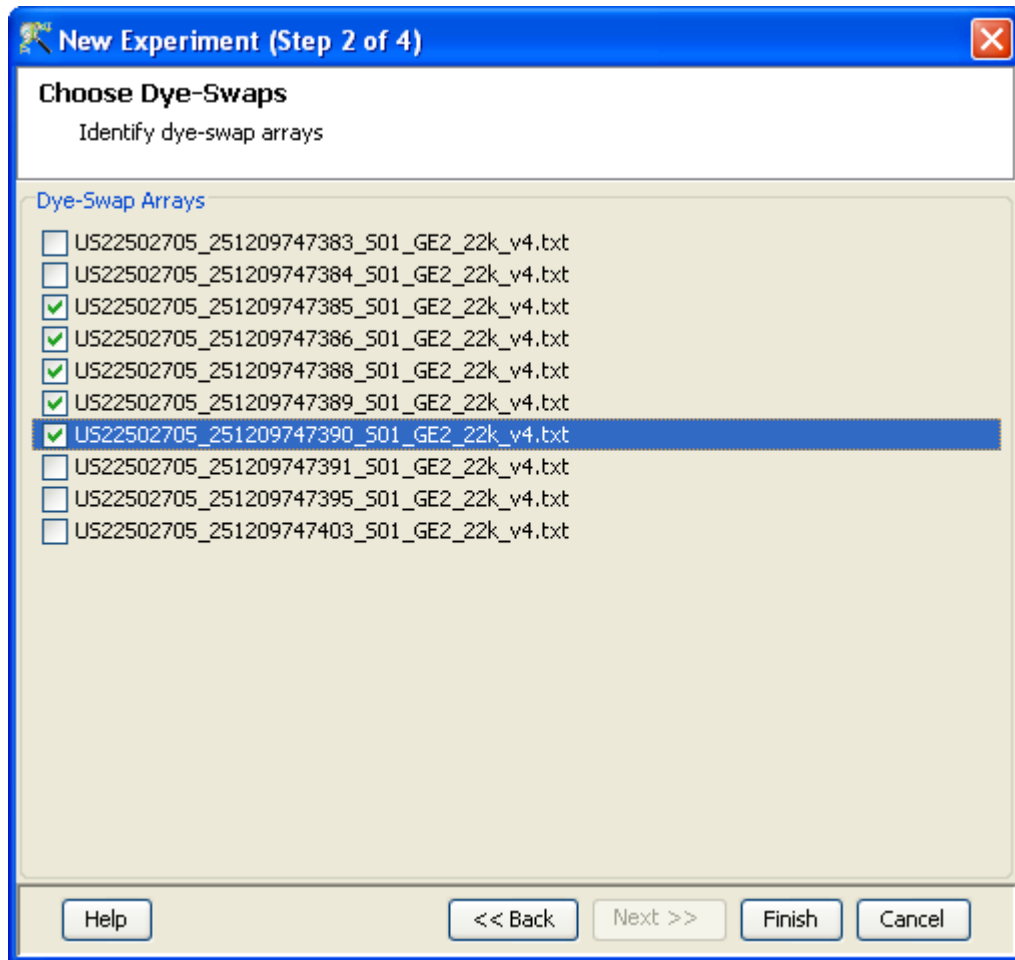


Figure 12.25: Choose Dye-Swaps

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click **Finish**. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

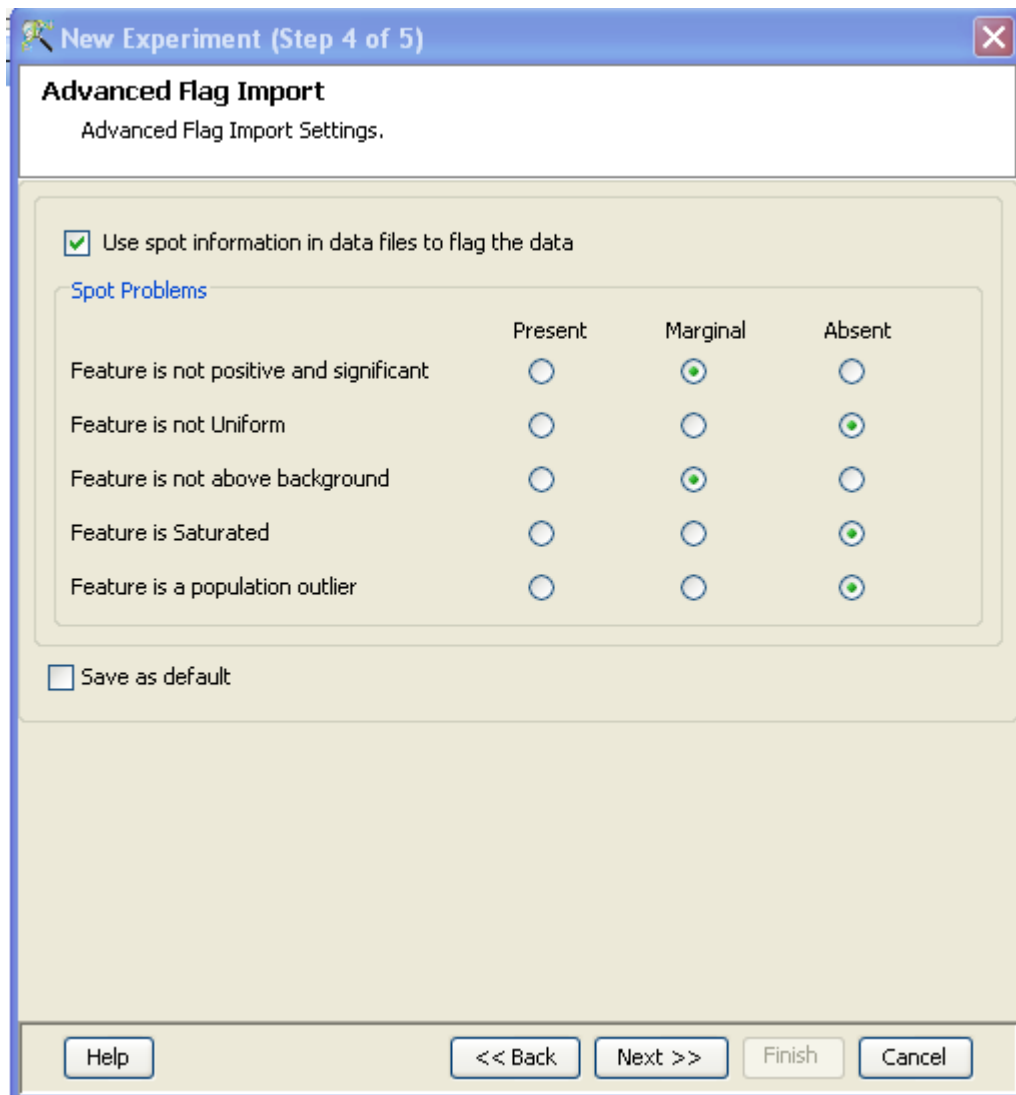


Figure 12.26: Advanced flag Import

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

For two colour, raw values are summarized for each channel separately and then log ratios are taken.

12.4.2 Quality Control

- **Quality Control on Samples:**

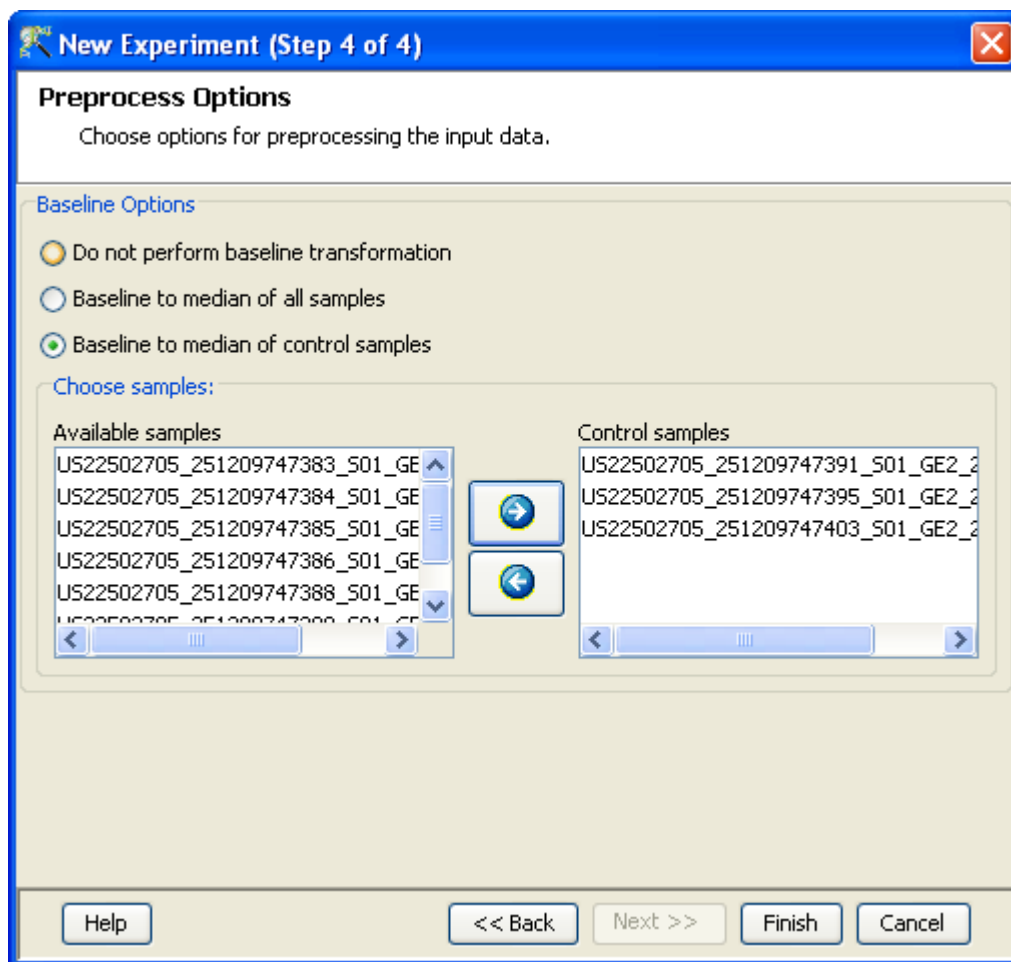


Figure 12.27: Preprocess Options

This view does not work with Agilent two colour files in .gpr format.

The view shows four tiled windows.

- Quality Metrics Report, Quality Metrics plot and Experiment Grouping tabs.
- PCA scores
- Legend

Figure 12.32 has the 4 tiled windows which reflect the QC on samples.

The metrics report include statistical results to help you evaluate the reproducibility and reliability of your microarray data.

The table shows the following:

More details on this can be obtained from the Agilent Feature Extraction Software(v9.5) Reference Guide, available from <http://chem.agilent.com>.

Quality controls Metrics Plot shows the QC metrics present in the QC report in the form of a plot.

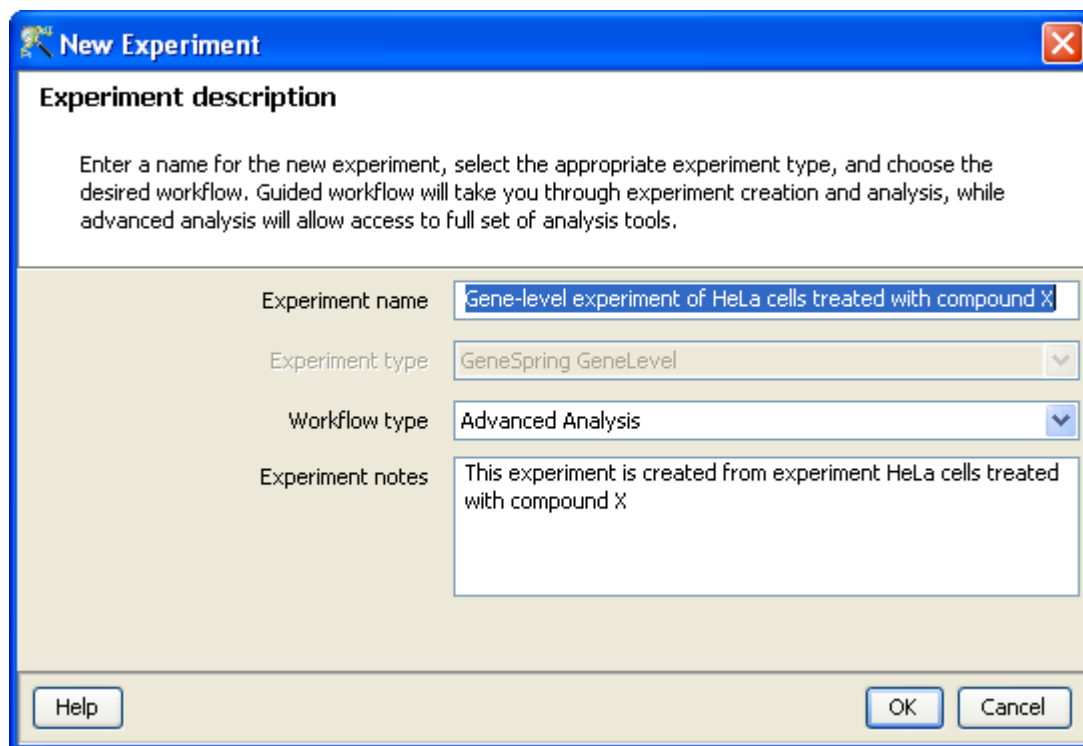


Figure 12.28: Gene Level Experiment Creation

Experiment grouping shows the parameters and parameter values for each sample.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click** → **Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The fourth window shows the legend of the active QC tab.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, summarization as well as baseline transformation is performed on the samples. Click on *OK* to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags:** In this step, the entities are filtered based on their flag values, the P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new experiment (Step 3 of 4) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

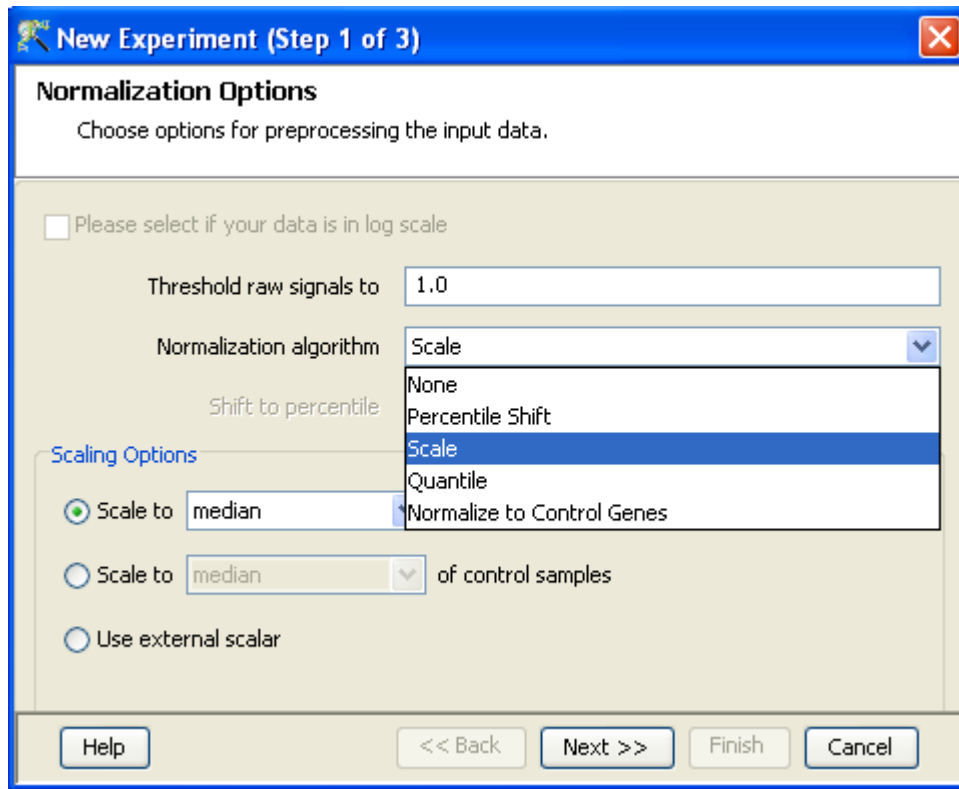


Figure 12.29: Gene Level Experiment Creation - Normalization Options

1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.
2. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected.
3. Step 2 of 4: This step is used to set the filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.
4. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window (See Figure 12.35).
5. Step 4 of 4: Click *Next* to annotate and save the entity list. (See Figure 12.36)

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on [Filter Probesets on Data Files](#)
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

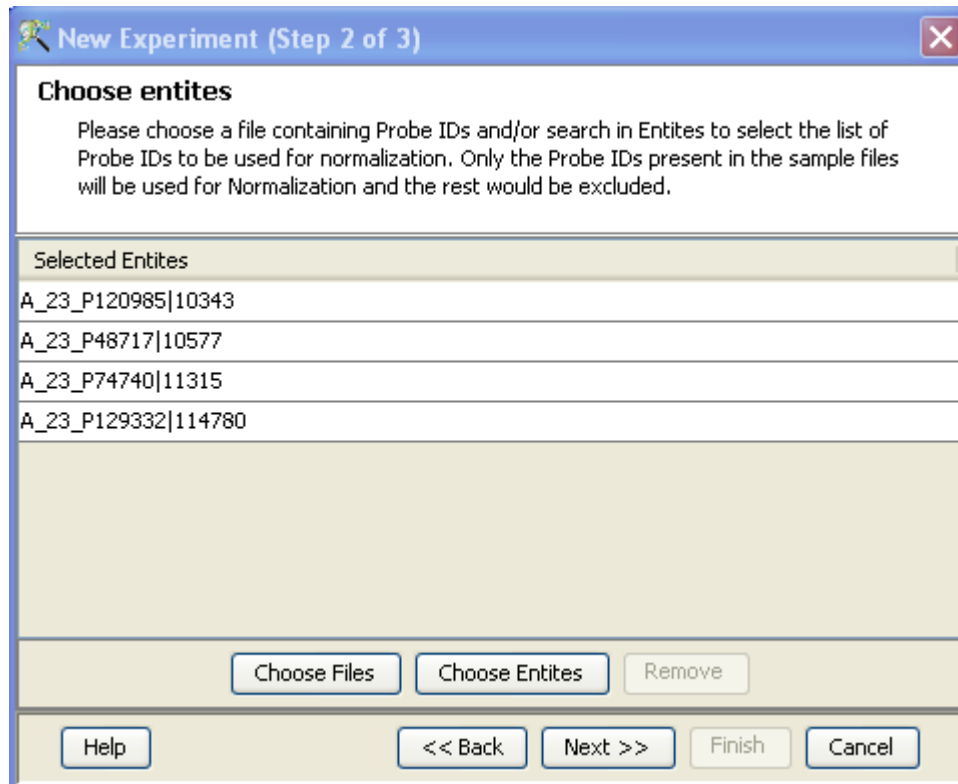


Figure 12.30: Gene Level Experiment Creation - Choose Entities

12.4.3 Analysis

- **Statistical Analysis**
For details refer to section [Statistical Analysis](#) in the advanced workflow.
- **Filter on Volcano Plot**
For details refer to section [Filter on Volcano Plot](#)
- **Fold Change**
For details refer to section [Fold Change](#)
- **Clustering**
For details refer to section [Clustering](#)
- **Find Similar Entities**
For details refer to section [Find Similar Entities](#)
- **Filter on Parameters**
For details refer to section [Filter on Parameters](#)
- **Principal Component Analysis**
For details refer to section [PCA](#)

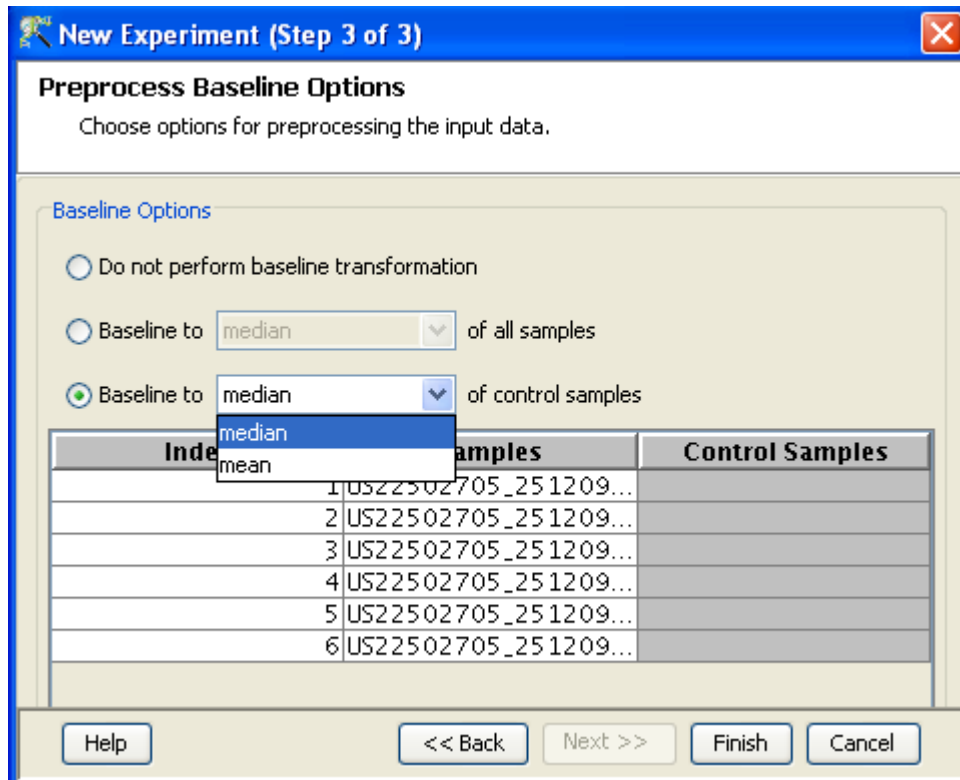


Figure 12.31: Gene Level Experiment Creation - Preprocess Baseline Options

12.4.4 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

12.4.5 Results

- **Gene Ontology (GO) analysis**
GO is discussed in a separate chapter called [Gene Ontology Analysis](#).
- **Gene Set Enrichment Analysis (GSEA)**
Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).
- **Gene Set Analysis (GSA)**
Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).
- **Pathway Analysis**
Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).

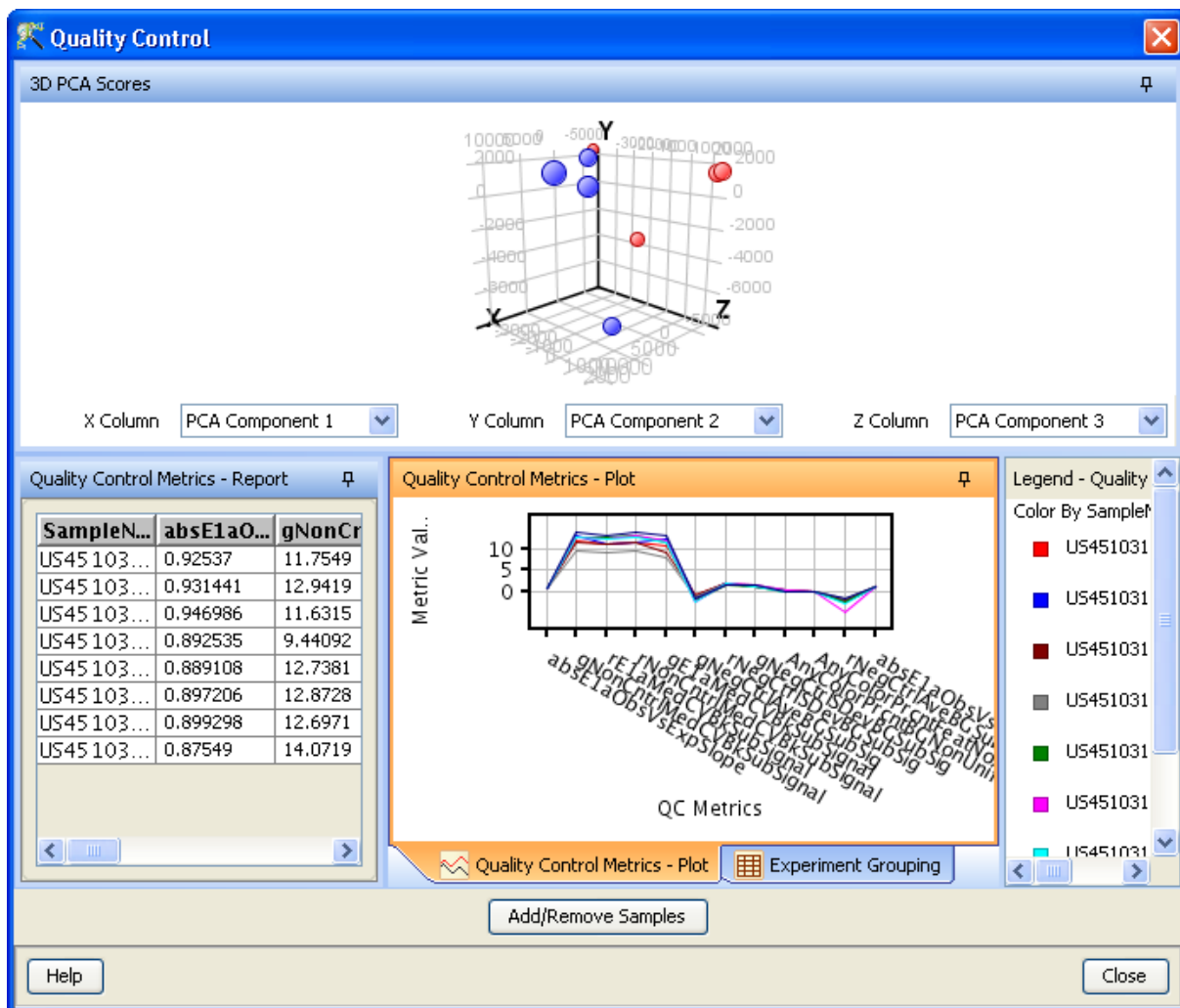


Figure 12.32: Quality Control

- **Find Similar Entity Lists**

This feature is discussed in a separate section called [Find Similar Entity Lists](#)

- **Find Significant Pathways**

This feature is discussed in a separate section called [Find Significant Pathways](#).

- **Launch IPA**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Import IPA Entity List**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Extract Interactions via NLP**

This feature is discussed in detail in the chapter [Pathway Analysis](#).

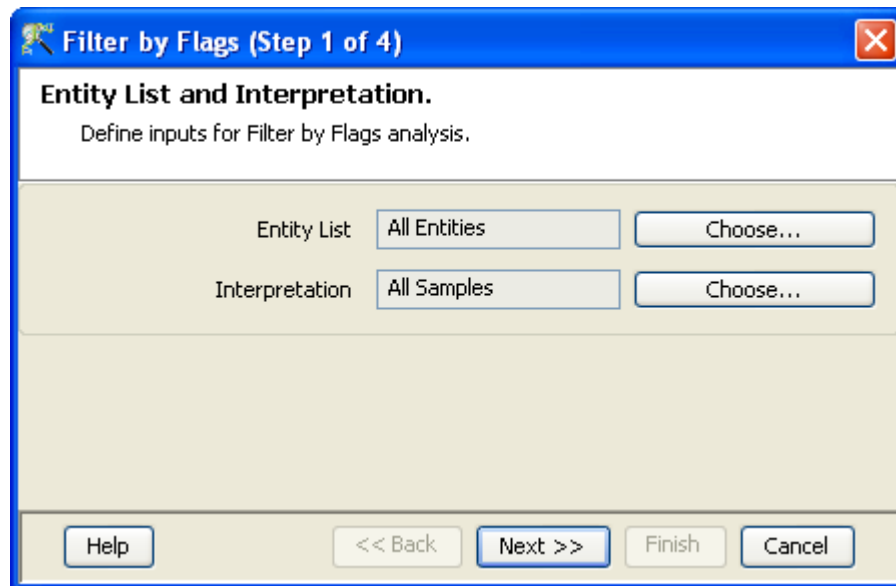


Figure 12.33: Entity list and Interpretation

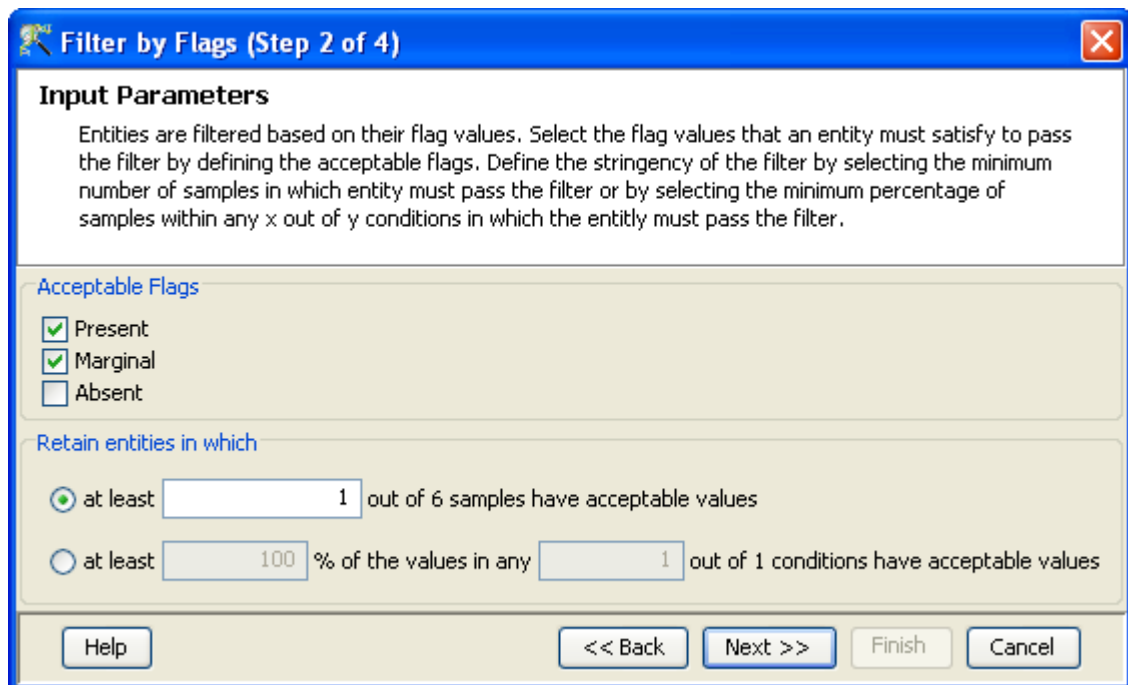


Figure 12.34: Input Parameters

Name of Metric	FE Stats Used	Description/Measures
absE1aObsVs ExpSlope	Abs(eQCObsVs ExpLRSlope)	Absolute of slope of fit for Observed vs. Expected E1a LogRatios
gNonCntrlMedCVBk SubSignal	gNonCntrlMedCVBk SubSignal	Median CV of replicated Non-Control probes: Green Bkgd-subtracted signals
rE1aMedCVBk SubSignal	reQCMedPrnt CVBGSubSig	Median CV of replicated E1a probes: Red Bkgd-subtracted signals
rNonCntrlMedCVBk SubSignal	rNonCntrlMedCVBk SubSignal	Median CV of replicated NonControl probes: Red Bkgd-subtracted signals
gE1aMedCVBk SubSignal	geQCMedPrnt CVBGSubSig	Median CV of replicated E1a probes: Green Bkgd-subtracted signals
gNegCtrlAve BGSubSig	gNegCtrlAve BGSubSig	Avg of NegControl Bkgd-subtracted signals (Green)
rNegCtrlAve BGSubSig	rNegCtrlAve BGSubSig	Avg of NegControl Bkgd-subtracted signals (Red)
gNegCtrlSDev BGSubSig	gNegCtrlSDev BGSubSig	StDev of NegControl Bkgd-subtracted signals (Green)
rNegCtrlSDevBGSUBSig	rNegCtrlSDevBGSUBSig	StDev of NegControl Bkgd-subtracted signals (Red)
AnyColorPrnt BGNonUnifOL	AnyColorPrnt BGNonUnifOL	Percentage of LocalBkgdRegions that are NonUnifOlr in either channel
AnyColorPrnt FeatNonUnifOL	AnyColorPrnt FeatNonUnifOL	Percentage of Features that are NonUnifOlr in either channel
absE1aObsVs ExpCorr	Abs(eQCObsVs ExpCorr)	Absolute of correlation of fit for Observed vs. Expected E1a LogRatios

Table 12.10: Quality Controls Metrics

12.4.6 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)
- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)
- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)
- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

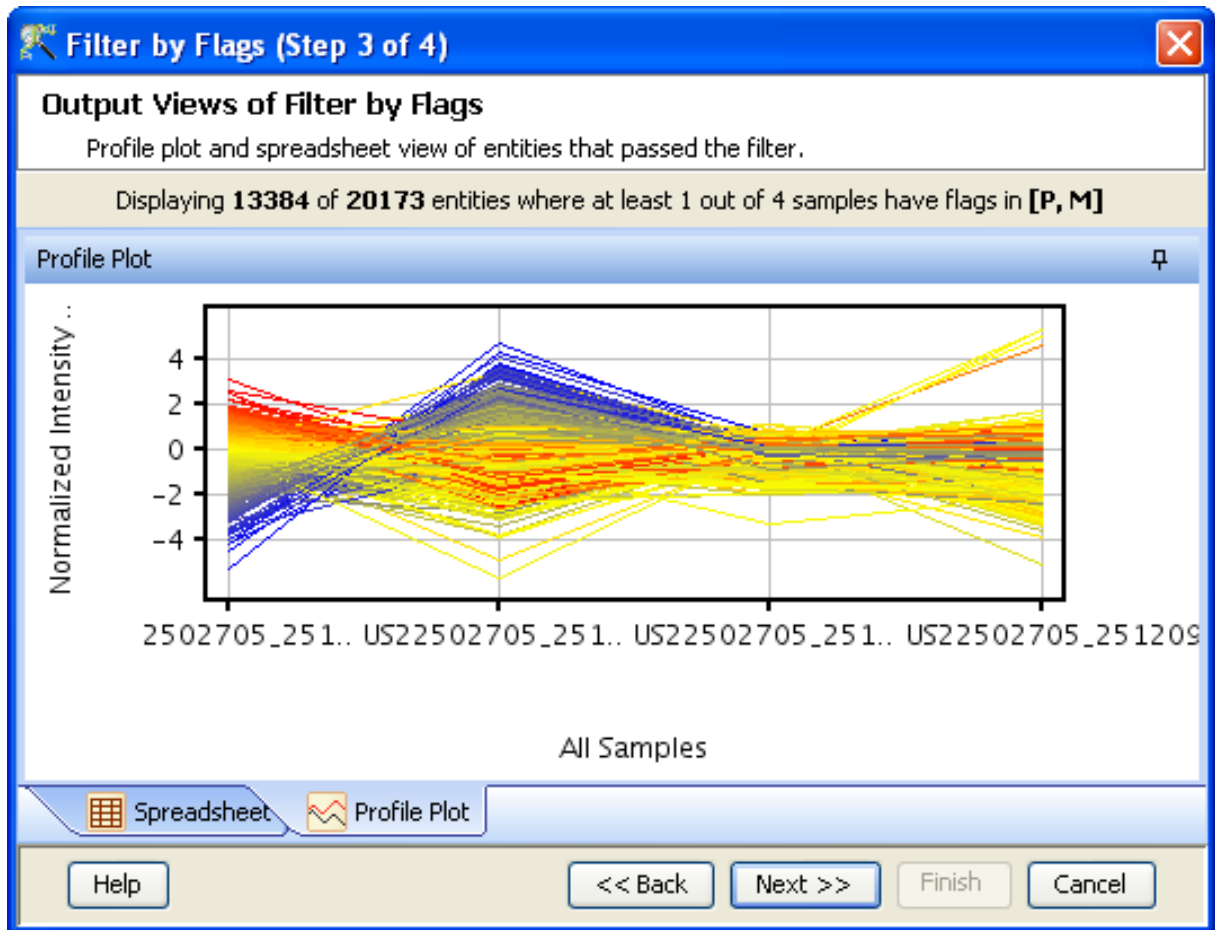


Figure 12.35: Output Views of Filter by Flags

12.5 Custom Agilent Arrays

The standard Agilent arrays can be analyzed using the Agilent Single and the Two Color Experiment types. In addition, **GeneSpring GX** also allows the user to analyze Custom Agilent arrays using Generic Single/Two Color Experiment types. To perform a Generic Single/Two Color analysis using Agilent arrays, the files can be an output from any FE, so long as they are in a tabular format. An annotation file is also required. Analysis through the Generic Two Color and Single Color workflows involves creation of a custom technology (Refer to chapters 15 and 16 on [Creating Technology](#) in Generic Single Color and [Creating Technology](#) in Generic Two Color) and specific markings of columns to perform GO, GSEA, or to view in Genome Browser. miRNA files can also be analyzed similarly.

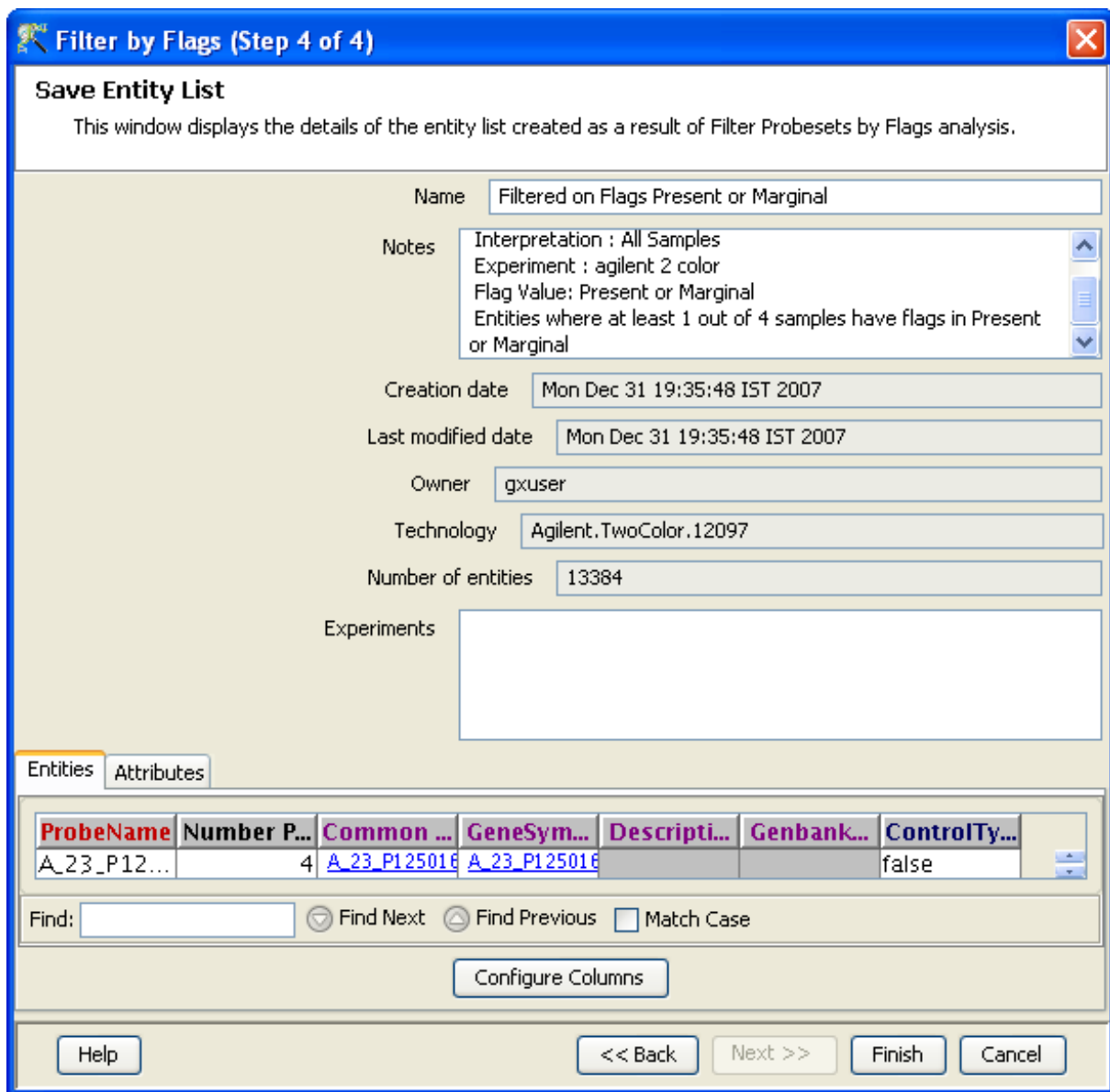


Figure 12.36: Save Entity List

Chapter 13

Analyzing Agilent miRNA Data

Micro RNAs or the miRNAs are small (22-25 nucleotides long), single-stranded, regulatory RNA molecules identified first in 1993. They are found in almost all of the life forms. Regulation is achieved by binding to regions of mRNA which share sequence complementarity with the miRNA. miRNAs affect the expression of genes involved in several physiological, developmental and pathological processes. Hence, expression studies of miRNA became important to understand their role in controlling biological and pathological processes. Advanced high throughput technologies like expression arrays enable us to study expression patterns of miRNA under given conditions. These studies can be correlated with their target gene expression studies.

GeneSpring GX supports all the Agilent miRNA microarray chip types. It supports data files obtained in text (.txt) format from Agilent Feature Extraction (FE) version 8.5 or 9.5.3. GeneView files are not supported.

13.1 Running the Agilent miRNA Workflow

Upon launching **GeneSpring GX**, the startup is displayed with 3 options.

- **Create new project**
- **Open existing project**
- **Open recent project**

Either a new project can be created or a previously generated project can be opened and re-analyzed. On selecting **Create new project**, a window appears in which details (Name of the project and Notes) can be recorded. **Open recent project** lists all the projects that were recently worked on and allows the user to select a project. After selecting any of the above 3 options, click on **OK** to proceed.

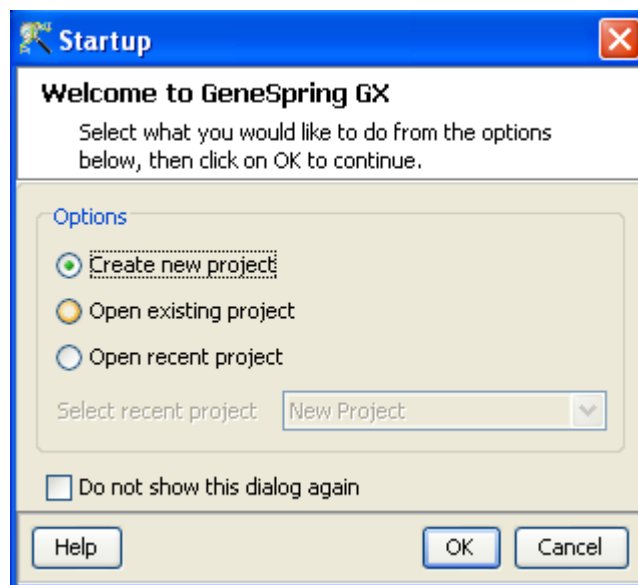


Figure 13.1: Welcome Screen

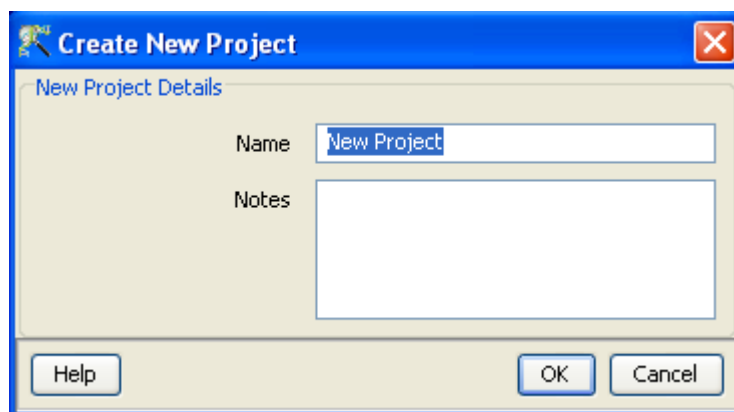


Figure 13.2: Create New project

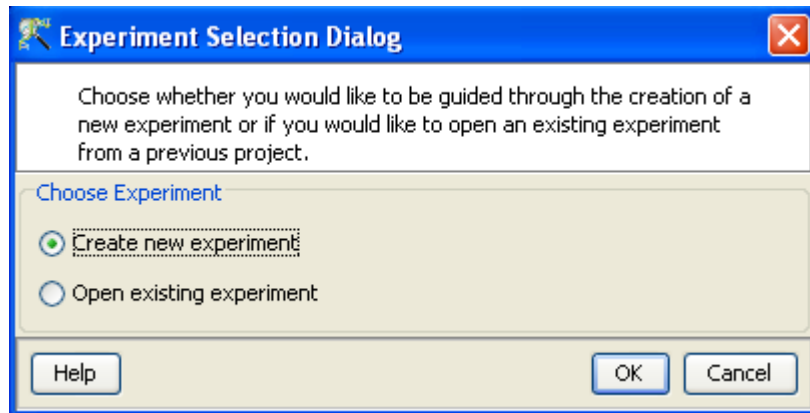


Figure 13.3: Experiment Selection

If **Create new project** is chosen, then an Experiment Selection dialog window appears with two options

1. **Create new experiment:** This allows the user to create a new experiment. (steps described below).
2. **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

Clicking on **Create new experiment** opens up a New Experiment dialog in which **Experiment name** can be assigned. The drop-down menu for the experiment type gives the user the option to choose between the multiple experiment types namely Affymetrix Expression, Affymetrix Exon Expression, Affymetrix Exon Splicing, Illumina Single Color, Agilent One Color, Agilent Two Color, Agilent miRNA, Generic Single Color, Generic Two Color, Pathway and RealTime-PCR experiment.

Next, the workflow type needs to be selected from the options provided below, based on the user convenience.

1. **Guided Workflow**
2. **Advanced Analysis Workflow**

Guided Workflow is primarily meant for a new user and is designed to assist the user through the creation and basic analysis of an experiment. Analysis involves default parameters which are not user configurable. However in **Advanced Analysis**, the parameters can be changed to suit individual requirements.

Upon selecting the workflow, a window opens with the following options:

1. Choose Files(s)

2. Choose Samples
3. Reorder
4. Remove

An experiment can be created using either the data files or else using samples. **GeneSpring GX** differentiates between a data file and a sample. A data file refers to the hybridization data obtained from a scanner. On the other hand, a sample is created within **GeneSpring GX**, when it associates the data files with its appropriate technology (See the section on [Technology](#)). Thus a sample created with one technology cannot be used in an experiment of another technology. These samples are stored in the system and can be used to create another experiment of the same technology via the **Choose Samples** option. For selecting data files and creating an experiment, click on the **Choose File(s)** button, navigate to the appropriate folder and select the files of interest. Click on **OK** to proceed.

Clicking on the **Choose Samples** button, opens a sample search wizard, with the following search conditions:

1. **Search field:** Requires one of the 6 following parameters- Creation date, Modified date, Name, Owner, Technology, Type can be used to perform the search.
2. **Condition:** Requires one of the 4 parameters- Equals, Starts with, Ends with and Includes Search value.
3. **Search Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment by clicking on **Add** button, or can be removed from the list using **Remove** button.

Files can either be removed or reordered during the data loading step using the **Remove** or **Reorder** button.

Figures [13.4](#) and [13.5](#), show the process of choosing experiment type and loading data.

GeneSpring GX creates the technology on the fly using user provided data identifiers. See figures [13.6](#), [13.7](#) and [13.8](#). Annotations from a file can be added at any time by going to **Annotations**→**Update Technology Annotations From file or Biological Genome**. For more details on technology creation in miRNA, refer to the section on [Technology creation on the fly](#). If an experiment has been created previously with the same technology, **GeneSpring GX** then directly proceeds with experiment creation.

Upon clicking **OK** in the **Load Data** window, the Agilent miRNA workflow appears. If the **Guided Workflow** option is chosen, the Guided Workflow wizard appears with the sequence of steps on the left

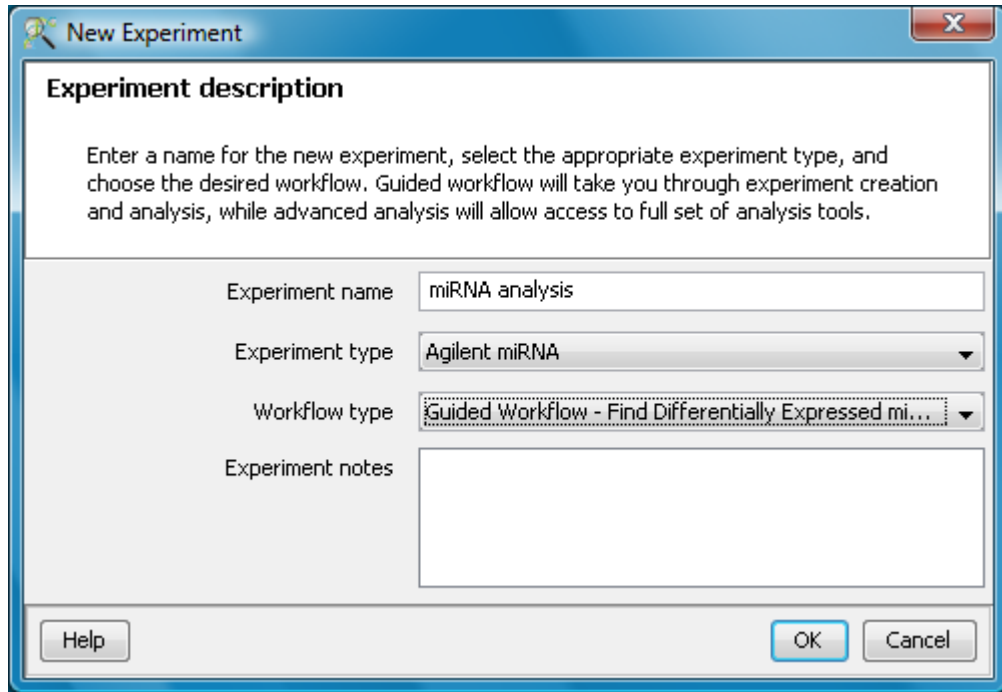


Figure 13.4: Experiment Selection

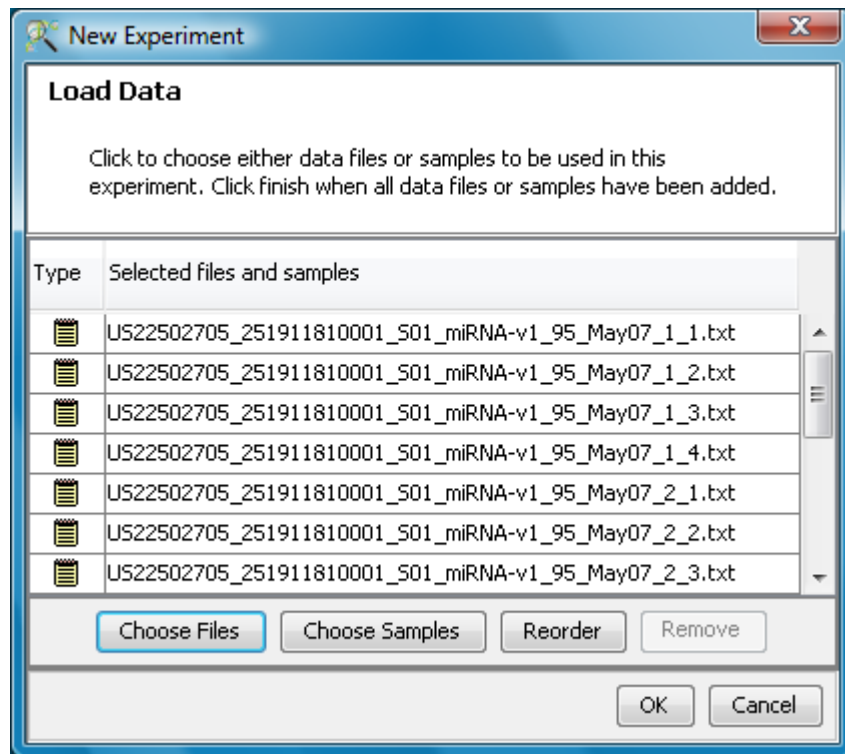


Figure 13.5: Load Data

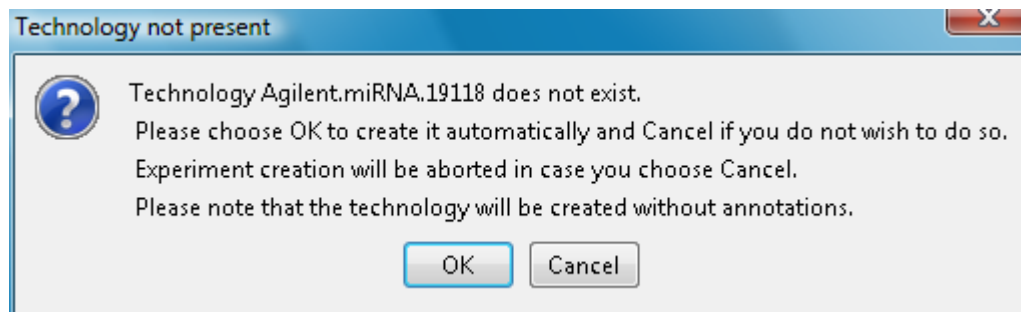


Figure 13.6: Technology Creation in miRNA

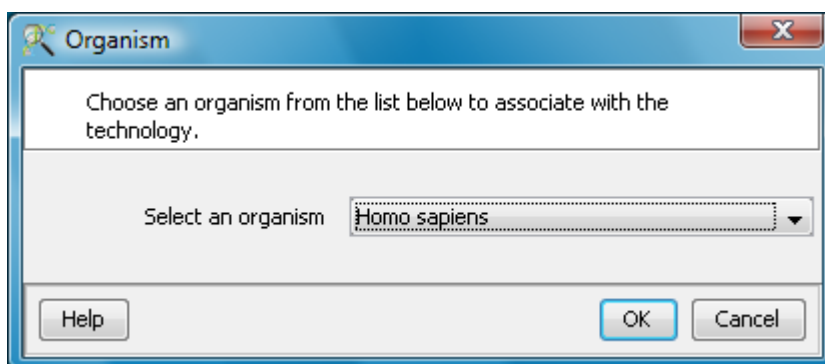


Figure 13.7: Selection of Organism

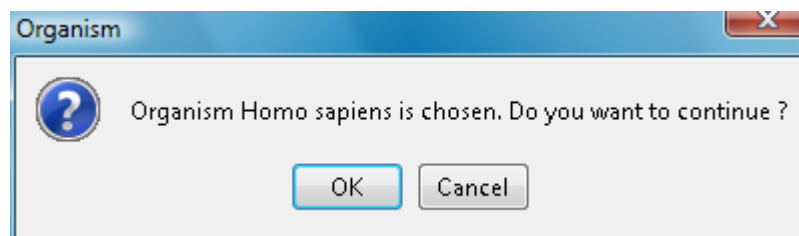


Figure 13.8: Confirmation Window

side highlighting the current step. The workflow allows the user to proceed in schematic fashion and does not allow the user to skip steps. If the **Advanced Analysis** has been chosen, then the step 2 of 4 of miRNA experiment creation wizard appears. For more details on experiment creation via **Advanced Workflow**, refer to the section on [Advanced Workflow](#).

13.1.1 Sample validation in GeneSpring GX 11.0

In **GeneSpring GX 11.0**, the AMADID field and the 'Grid.Date' field are both considered as unique identifiers for miRNA arrays. The 'Grid Date' field gives information on the version of the design file that was used to extract the data during sample creation. This means that even if the AMADID is same for the samples (for example, 19118), if they were created using a different design file, then they will not be taken together for experiment creation.

In **GeneSpring GX 10.0**, only the AMADID field was the unique identifier. Hence, while moving from **GeneSpring GX 10.0** to **GeneSpring GX 11.0**, it is recommended that users recreate miRNA experiments in GX11. To recreate, download the samples from the GX 10.0 experiment (right click on the 'Samples' folder in the experiment and choose 'Download Samples'). During this process, the Grid.Date field will be extracted and stored within. Create a new experiment by using these as 'files' and not as 'samples', in GX 11.0.

13.2 Data Processing

- **File formats:** The data files should in text (.txt) format and obtained from Agilent Feature Extraction (FE) 8.5 and 9.5.3. **GeneSpring GX** supports the full file format and does not support the GeneView format files.
- **Raw Signal Values:** The term "raw" signal values refer to the linear data after thresholding and summarization. Summarization is done by taking the geometric mean in **GeneSpring GX** .
- **Normalized Signal Values:** "Normalized" value is the value generated after log transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and Baseline Transformation.
- **Treatment of on-chip replicates:** The signal value of a probeset is the geometric mean of all its probes.
- **Flag values:** The flag value of a particular probeset is dependant on the flag values of the probes in it. The 'gIsGeneDetected' is taken as flag column and a value of 0 is considered as Absent and 1 is considered as Present.
- **Treatment of Control probes:** The control probes are included while performing normalization.
- **Empty Cells:** Not Applicable.

- **Sequence of events:** The sequence of events involved in the processing of the data files is: Thresholding→S Transformation→Normalization→Baseline Transformation.

13.3 Guided Workflow steps

13.3.1 Summary Report (Step 1 of 8)



The Summary report displays the summary view of the created experiment. It shows a Box Whisker plot, with the samples on the X-axis and the Log Normalized Expression values on the Y axis. An information message on the top of the wizard shows the number of samples in the file and the sample processing details. By default, the *Guided Workflow* does a thresholding of the signal values to 1. It then normalizes the data to 75th percentile and does not perform baseline transformation. If the number of samples are more than 30, they are only represented in a tabular column. On clicking the **Next** button it will proceed to the next step and on clicking **Finish**, an entity list will be created on which analysis can be done. By placing the cursor on the screen and selecting by dragging on a particular probe, the probe in the selected sample as well as those present in the other samples are displayed in green. Figure 13.9 shows the Summary report with box-whisker plot.

Note: In the *Guided Workflow*, these default parameters cannot be changed. To choose different parameters use *Advanced Analysis*.

13.3.2 Experiment Grouping (Step 2 of 8)

On clicking **Next**, the *Experiment Grouping* window appears which is the 2nd step in the **Guided Workflow**. It requires parameter values to be defined to group samples. Samples with same parameter values are treated as replicates. To assign parameter values, click on the **Add parameter** button. Parameter values can be assigned by first selecting the desired samples and assigning the corresponding parameter value. For removing any value, select the sample and click on **Clear**. Press **OK** to proceed. Although any number of parameters can be added, only the first two will be used for analysis in the **Guided Workflow**. The other parameters can be used in the **Advanced Analysis**.

Note: The *Guided Workflow* does not proceed further without grouping information.

Experimental parameters can also be loaded externally by clicking on Load experiment parameters from file  icon button. The file containing the *Experiment Grouping* information should be a tab or comma separated text file. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file should contain a

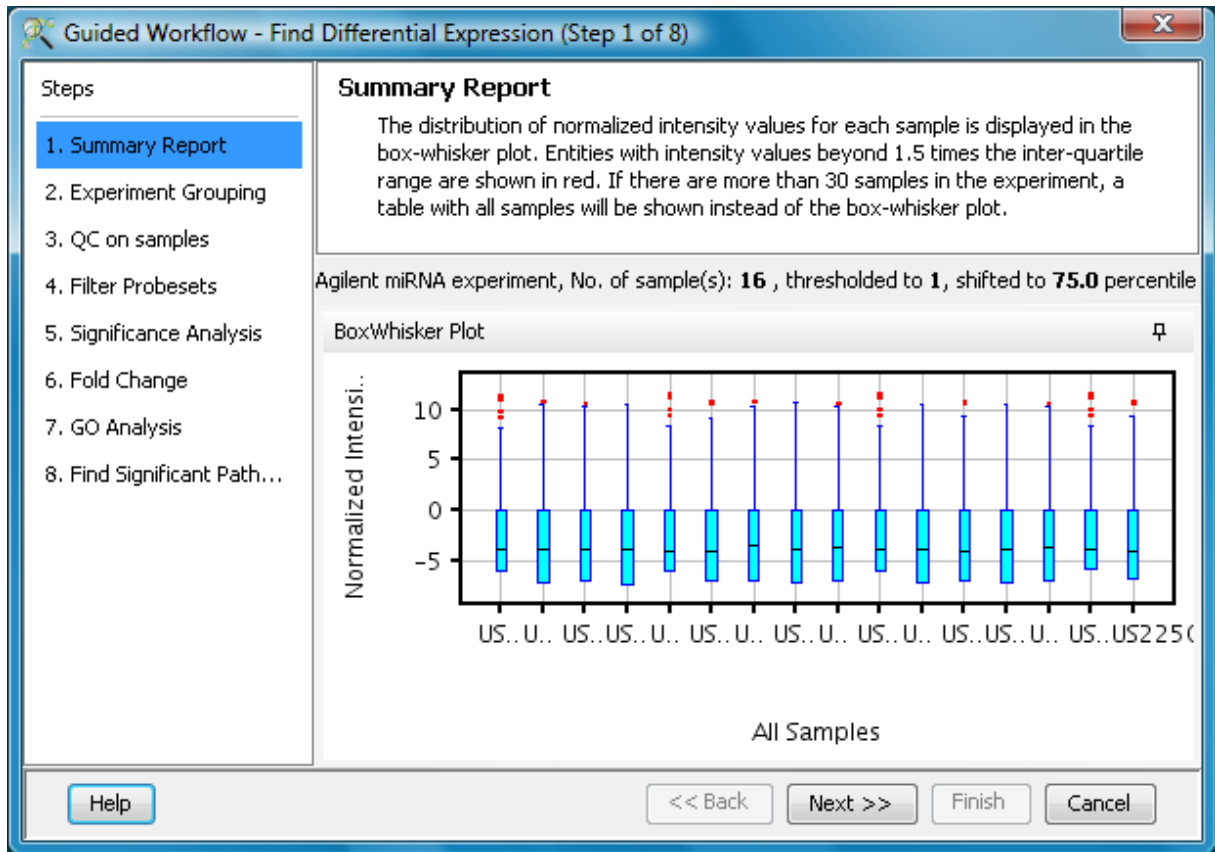



Figure 13.9: Summary Report




column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated text file.

Sample genotype dosage

```
A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50
```

Reading this tab file generates new columns corresponding to each factor.

The current set of experiment parameters can also be saved to a local directory as a tab separated or comma separated text file by clicking on the Save experiment parameters to file  icon button. These saved parameters can then be imported and used for future analysis. In case of multiple parameters, the

individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click → *Properties* → *Columns* option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click → *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

Note: The *Guided Workflow* by default creates averaged and unaveraged interpretations based on parameters and conditions. It takes average interpretation for analysis in the guided wizard.

Windows for Experiment Grouping and Parameter Editing are shown in Figures [13.10](#) and [13.11](#) respectively.

13.3.3 Quality Control (QC) (Step 3 of 8)

The 3rd step in the Guided workflow is the QC on samples which is displayed in the form of four tiled windows. They are as follows:

- Quality Controls Metrics- Report and Experiment grouping tabs
- Quality Controls Metrics- Plot
- 3D PCA Scores.
- Legend

QC generates four tiled windows as seen in figure [13.12](#).

The *Experiment Grouping* tab shows the grouping information specified in the previous step.

The metrics report helps the user evaluate the reproducibility and reliability of the microarray data. The quality metrics scores are obtained directly from the sample file. A brief description is given below:

- Additive error (AddErrorEstimateGreen): measures on feature background noise. Should be <5, 5~12 is concerning, >12 is bad

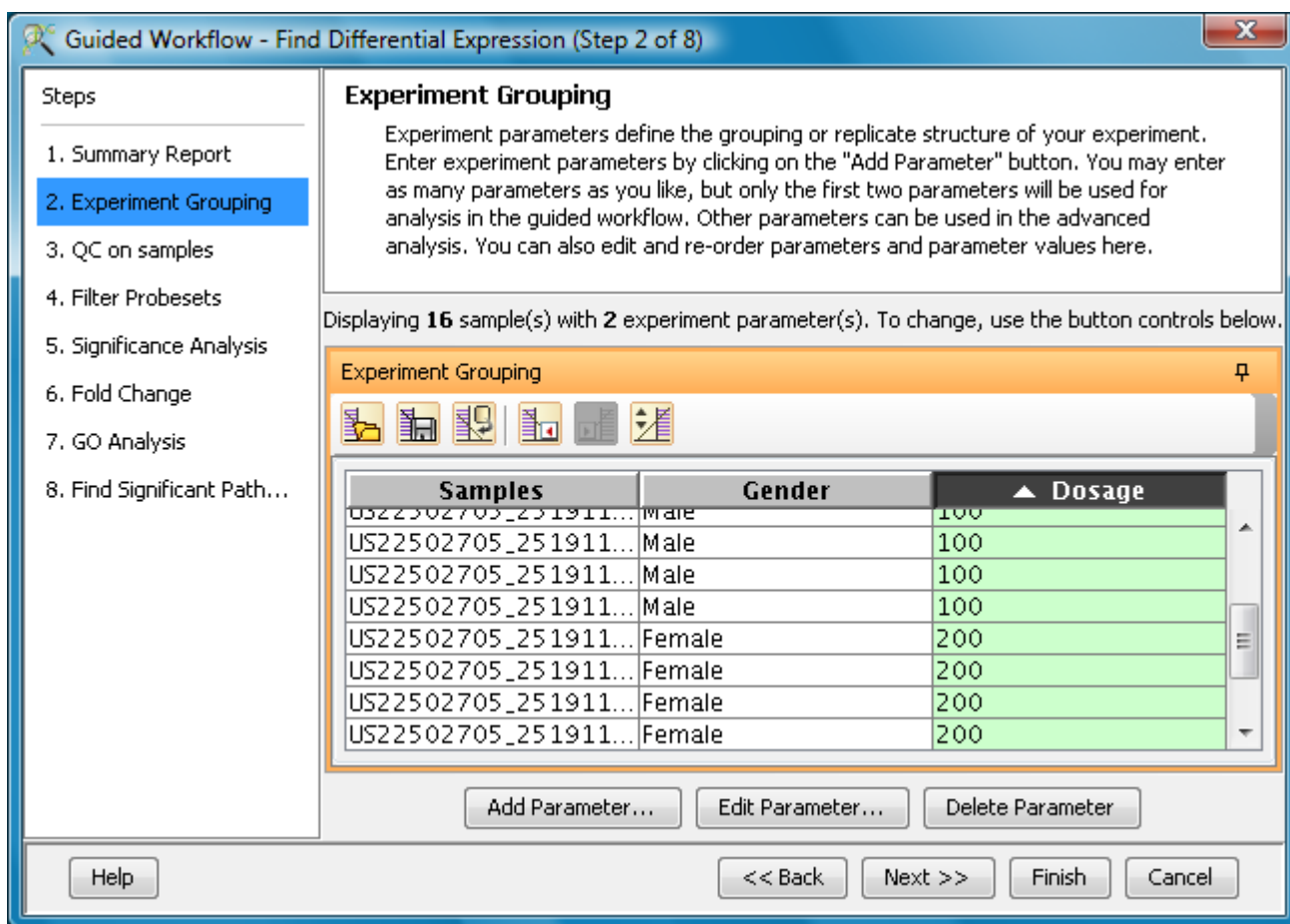


Figure 13.10: Experiment Grouping

- % Feature Population Outlier (AnyColorPrctFeatPopnOL): Measures % of features that are called population outliers (and therefore excluded from analysis) Should be less than 8%, >~15% is bad
- NonControl %CV of BGsubtracted Signal (gNonCtrlMedPrctCVBGSig): Measures uniformity of signals across feature replicates Should be <10%, >~15% is bad, -1 is bad
- 75% ile Total Gene Signal (gTotalSignal75pctile): Measures overall intensity of non control probes. This metric is HIGHLY sample dependant, but should be consistent for well behaving samples of similar type.

More details on this can be obtained from the Agilent Feature Extraction Software(v9.5) Reference Guide, available from <http://chem.agilent.com>.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from

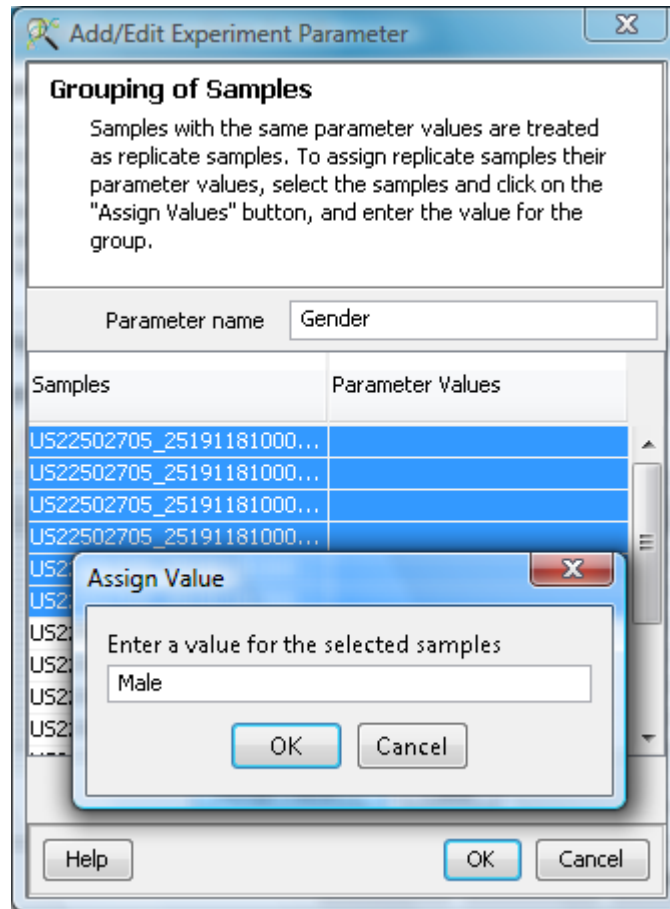


Figure 13.11: Add/Edit Parameters

arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The *Add/Remove* samples allows the user to remove the unsatisfactory samples and to add the samples back if required. Whenever samples are removed or added back, normalization is performed again. Click on **OK** to proceed.

The fourth window shows the legend of the active QC tab.

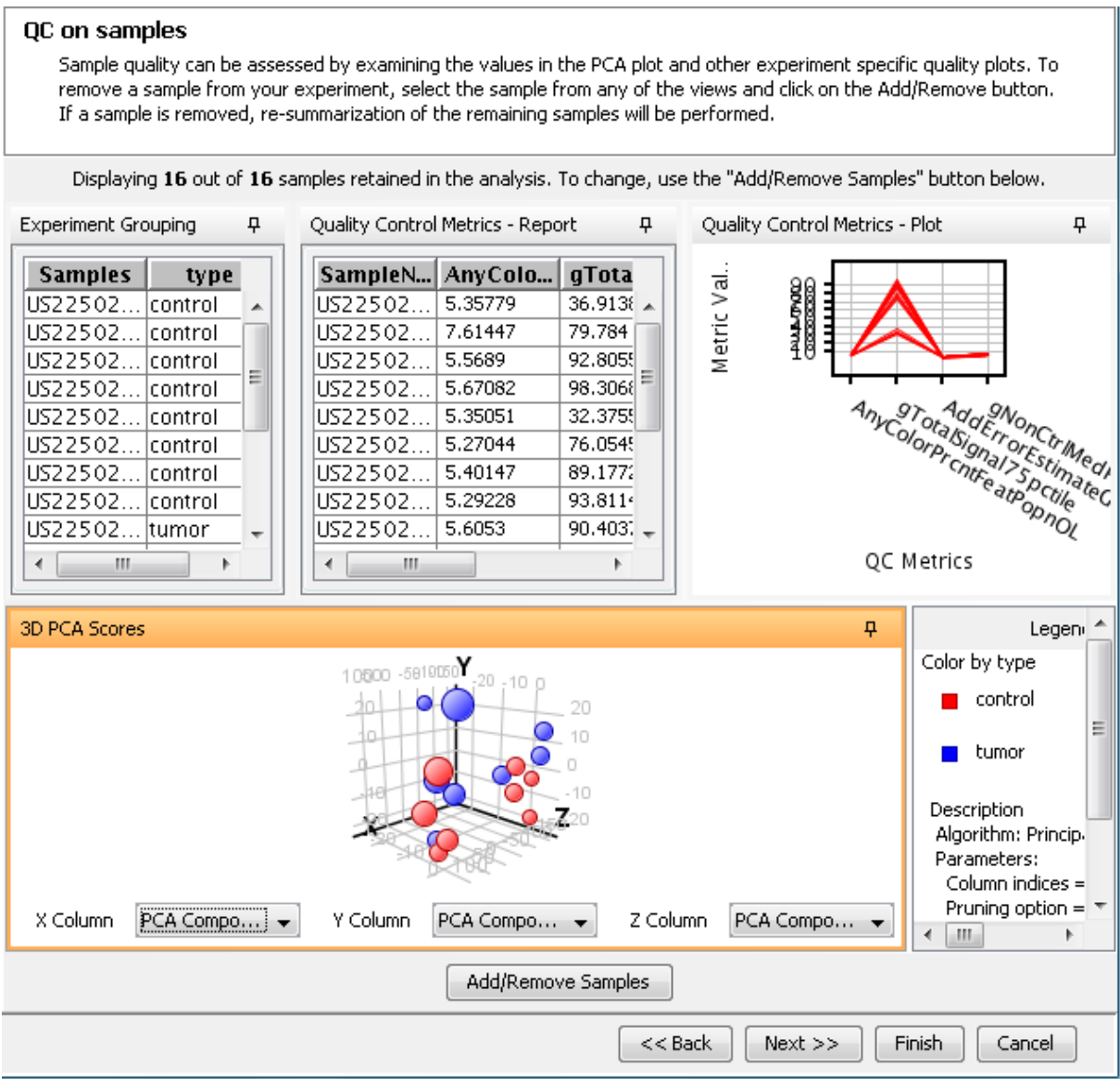


Figure 13.12: Quality Control on Samples

13.3.4 Filter probesets (Step 4 of 8)

In this step, the entities are filtered based on their flag values $P(\textit{present})$ and $A(\textit{absent})$. Information pertaining to the flags is present in the data file. **GeneSpring GX** considers the "gIsGeneDetected" as the flag column and marks entities having '0' as *Absent* and '1' as *Present*. Only entities having the present flag in at least 1 sample are displayed in the profile plot. The selection can be changed using *Rerun Filter* option. The plot is generated using the normalized signal values and samples grouped by the active interpretation. Options to customize the plot can be accessed via the Right-click menu. An *Entity List*, corresponding to this filtered list, will be generated and saved in the Navigator window. The Navigator window can be viewed after exiting from *Guided Workflow*. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the information corresponding to the selected entity. Newer annotations can be added and existing ones removed using the *Configure Columns* button. An additional tab in the *Entity Inspector* shows the raw and normalized values for that entity. A plot which shows the distribution of the normalized intensity values of that entity over the current interpretation is present as a tab in the same window. The cutoff for filtering can be changed using the *Rerun Filter* button. Newer Entity lists will be generated with each run of the filter and saved in the Navigator. The information message on the top shows the number of entities satisfying the flag values.

Figures 13.14 and 13.13 are displaying the profile plot obtained in situations having single and two parameters.

13.3.5 Significance Analysis (Step 5 of 8)

Depending upon the experimental grouping, **GeneSpring GX** performs either T-test or ANOVA. The tables below describe broadly the type of statistical test performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I*, has 2 groups, the normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 13.1: Sample Grouping and Significance Tests I

- **Example Sample Grouping II:** In this example, only one group, the tumor, is present. T-test against zero will be performed here.

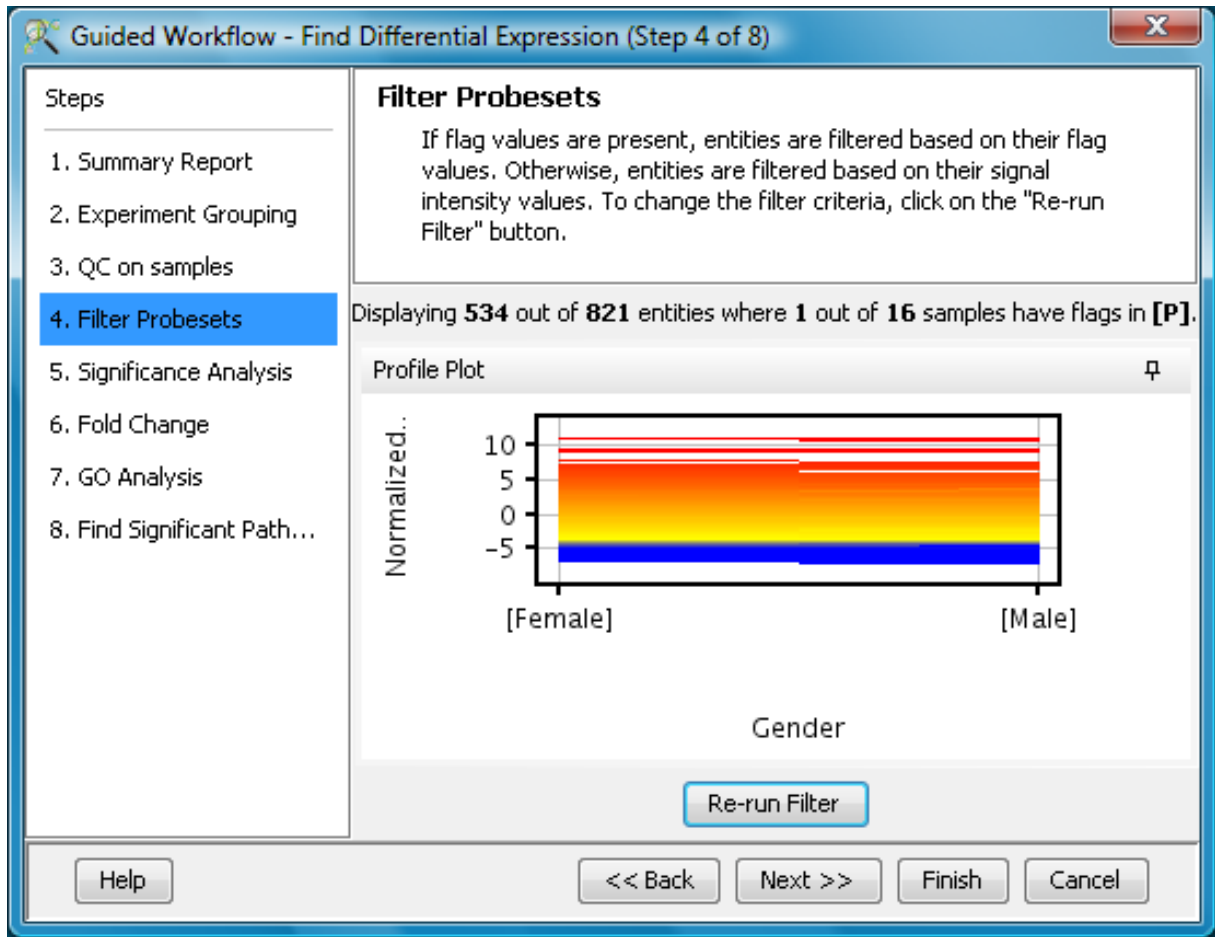


Figure 13.13: Filter Probesets-Single Parameter

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 13.2: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When 3 groups are present (normal, tumor1 and tumor2) and one of the groups (tumor2 in this case) does not have replicates, statistical analysis cannot be performed. However if the condition tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.
- **Example Sample Grouping IV:** When there are 3 groups within an interpretation, One-way ANOVA will be performed.

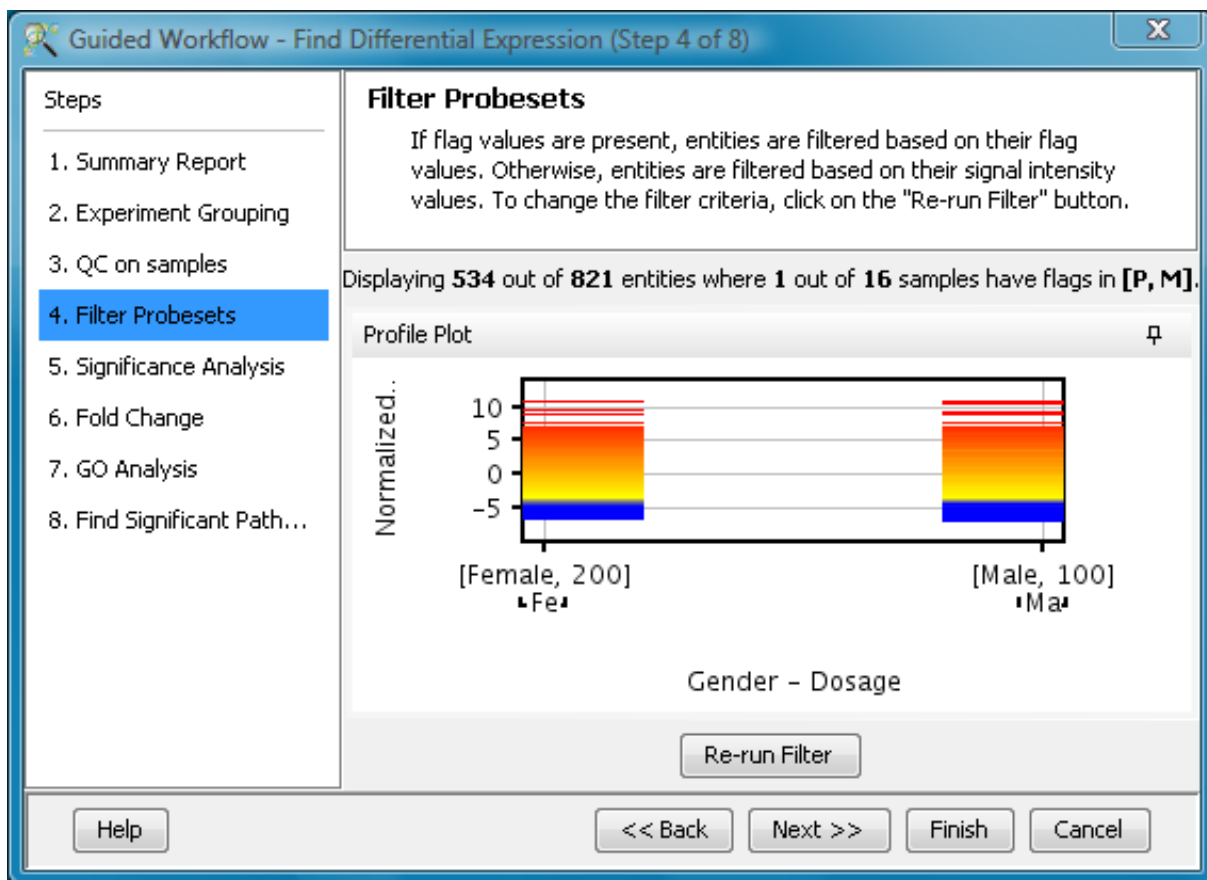


Figure 13.14: Filter Probesets-Two Parameters

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 13.3: Sample Grouping and Significance Tests III

- **Example Sample Grouping V:** This table shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.
- **Example Sample Grouping VI:** In this table, a two-way ANOVA will be performed.
- **Example Sample Grouping VII:** In the example below, a two-way ANOVA will be performed and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min,

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 13.4: Sample Grouping and Significance Tests IV

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 13.5: Sample Grouping and Significance Tests V

Tumor/30min, Tumor/50min), which is the same as the number of samples. The p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Statistical Tests: T-test and ANOVA

- **T-test: T-test unpaired** is chosen as a test of choice with a kind of experimental grouping shown in Table 1. Upon completion of T-test the results are displayed as three tiled windows.
 - A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *Regulation*.
 - *Differential expression analysis report* mentioning the Test description i.e. test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic or Permutative*).

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

- **Analysis of variance(ANOVA):** ANOVA is chosen as a test of choice under the experimental grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 13.6: Sample Grouping and Significance Tests VI

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumor	10 min
S5	Tumor	30 min
S6	Tumor	50 min

Table 13.7: Sample Grouping and Significance Tests VII

- A *p-value table* consisting of probe names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.
- *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and p-value computation type (*Asymptotic or Permutative*).
- *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

Special case: In situations when samples are not associated with at least one possible permutation of conditions (like Normal at 50 min and Tumor at 10 min mentioned above), no p-value can be computed and the **Guided Workflow** directly proceeds to **GO analysis**.

13.3.6 Fold-change (Step 6 of 8)

Fold change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions, Condition 1 and Condition 2. The ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 1/Condition 2). Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance

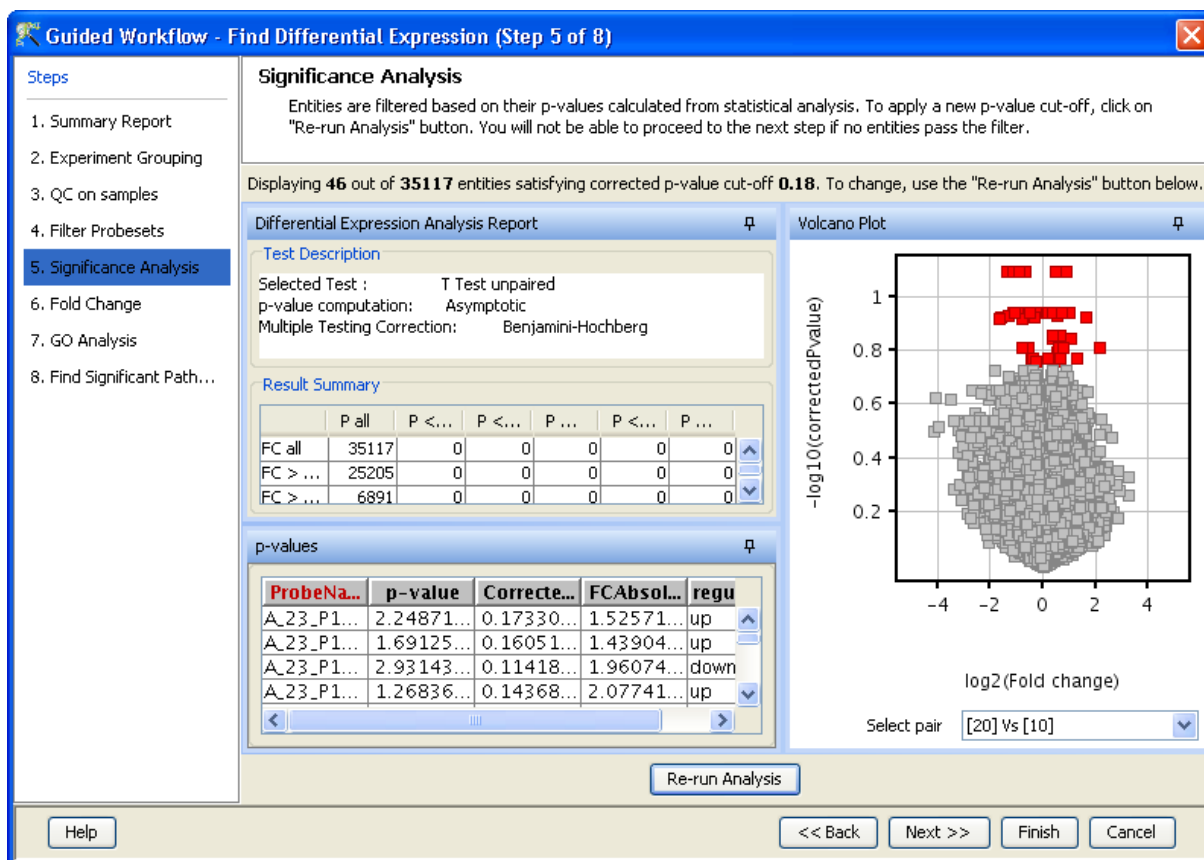


Figure 13.15: Significance Analysis-T Test

analysis are passed on for the fold change analysis. The wizard shows a table consisting of 3 columns: Probe Names, Fold change value and regulation (up or down). The regulation column depicts which one of the groups has greater or lower intensity values wrt other group. The cut off can be changed using **Re-run Filter**. The default cut off is set at 2.0 fold. So it shows all the entities which have fold change values greater than or equal to 2. The fold change value can be manipulated by either using the sliding bar (goes up to a maximum of 10.0) or by typing in the value and pressing Enter. Fold change values cannot be less than 1. A profile plot is also generated. Upregulated entities are shown in red. The color can be changed using the Right-click→*Properties* option. Double click on any entity in the plot shows the *Entity Inspector* giving the annotations corresponding to the selected entity. An entity list will be created corresponding to entities which satisfied the cutoff in the experiment Navigator.

Note: Fold Change step is skipped and the *Guided Workflow* proceeds to the *GO Analysis* in case of experiments having 2 parameters.

Fold Change view with the spreadsheet and the profile plot is shown in figure 13.17.

On clicking **Next**, the tool prompts the user that the database for the organism is not found and gives the option of downloading the same. See Figure 13.18. This refers to the TargetScan data for that particular

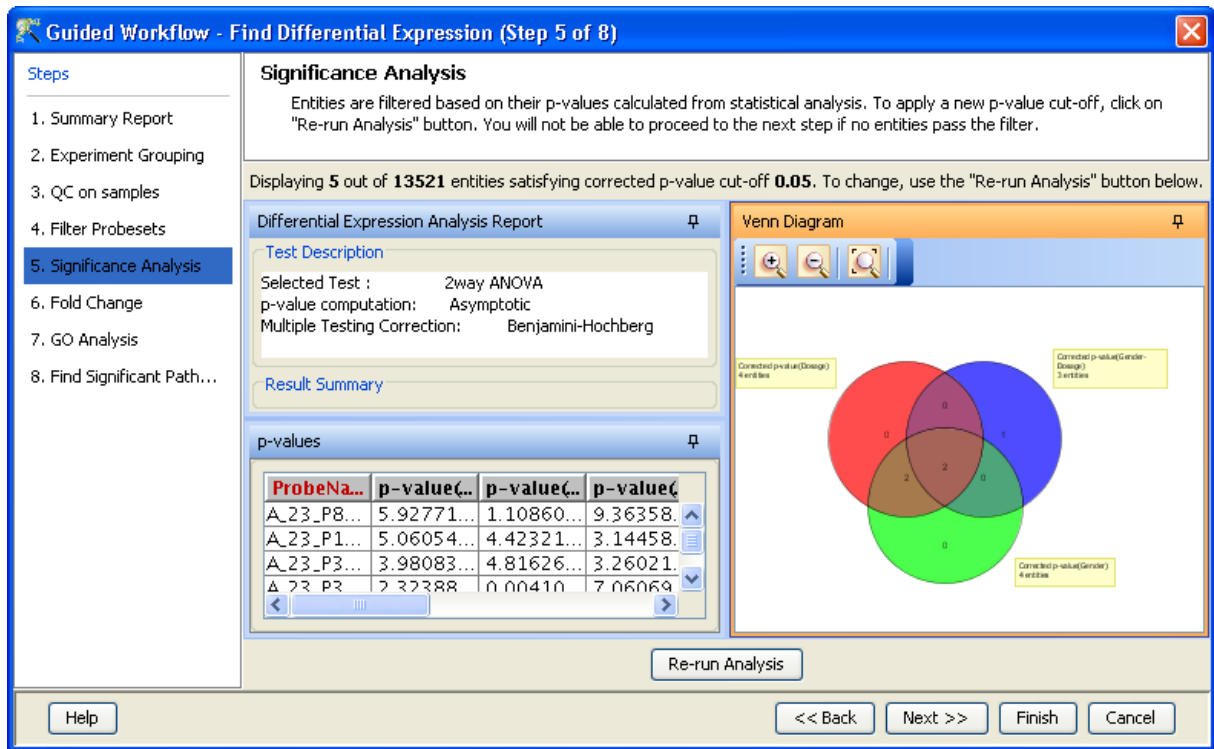


Figure 13.16: Significance Analysis-Anova

organism and it contains the mapping information for all the miRNA to their respective mRNA targets. **GeneSpring GX** uses the TargetScan database(Version-4.2) to predict the targets for the analyzed set of miRNA entities (See Section 13.4.6). This is essential for finding the genes that could be affected by the differentially expressed miRNAs. This tool uses the organism's TargetScan data to arrive at the mapping information for the entities in the entity list that is created as a result of fold Change. The default value taken for performing the TargetScan method in **Guided Workflow** is 50th percentile and the database used is the conserved database.

After the miRNAs are mapped to their respective genes, **GeneSpring GX** helps the user to find out their functions as well as the pathways in which these genes are involved via GO and pathway analysis. Both these analyses require specific annotation columns which are not present in the miRNA technology. Hence the tool prompts the user that the biological genome does not exist for that organism and gives the option of downloading the same. See Figure 13.19. Biological Genome is the term used for the collective set of annotations for a particular organism that can be built in **GeneSpring GX** and is essential in performing analysis such as GO Analysis, Genome Browser, Pathway etc. For more information on the same, refer to the section on [Biological Genome](#).

13.3.7 Gene Ontology Analysis (Step 7 of 8)

The *GO Consortium* maintains a database of controlled vocabularies for the description of molecular function, biological process and cellular location of gene products. The GO terms are displayed in the

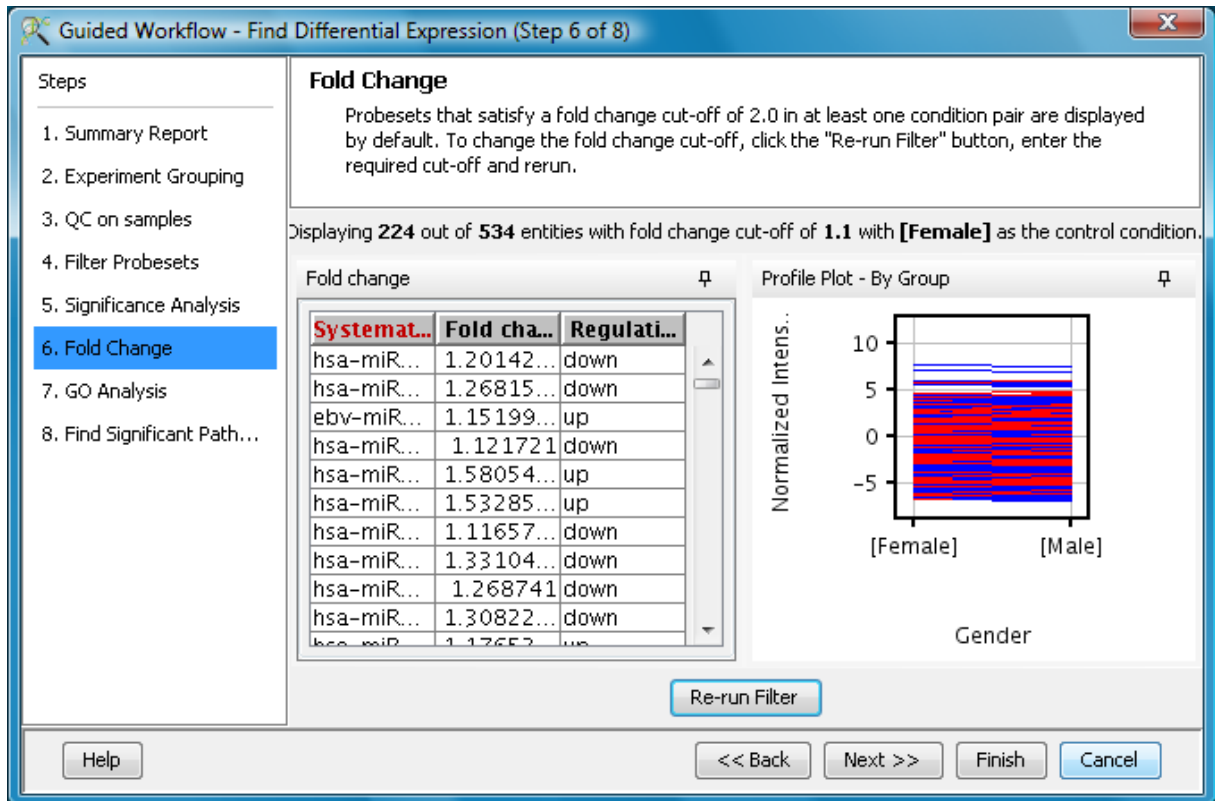


Figure 13.17: Fold Change

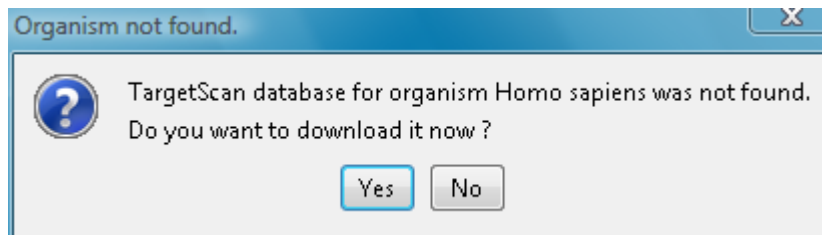


Figure 13.18: TargetScan Database Download

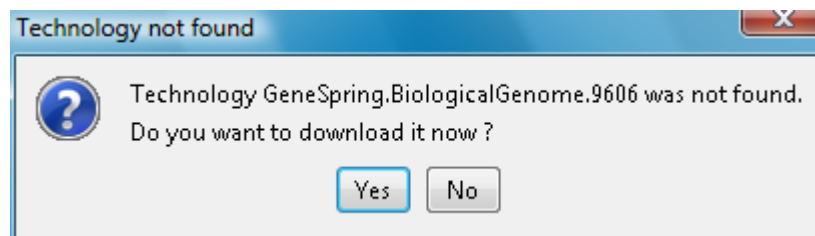


Figure 13.19: Biological Genome Download

Gene Ontology column with associated *Gene Ontology Accession* numbers. A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. Since the Gene Ontology is a Directed Acyclic Graph (DAG), GO terms can be derived from one or more parent terms. The Gene Ontology classification system is used to build ontologies. All the entities with the same GO classification are grouped into the same gene list.

The GO analysis wizard shows two tabs comprising of a spreadsheet and a *GO tree*. The *GO Spreadsheet* shows the *GO Accession* and *GO terms* of the selected genes. For each GO term, it shows the number of genes in the selection; and the number of genes in total, along with their percentages. Note that this view is independent of the dataset, is not linked to the master dataset and cannot be lassoed. Thus selection is disabled on this view. However, the data can be exported and views if required from the right-click. The p-value for individual GO terms, also known as the enrichment score, signifies the relative importance or significance of the GO term among the genes in the selection compared the genes in the whole dataset. The default p-value cut-off is set at 0.1 and can be changed to any value between 0 and 1.0. The GO terms that satisfy the cut-off are collected and the all genes contributing to any significant GO term are identified and displayed in the GO analysis results.

The GO tree view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. This GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of genes in the dataset that corresponds to the selected GO term(s). The selection operation is detailed below.

When the GO tree is launched at the beginning of GO analysis, the GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others are shown in black. Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off. The search button is also provided on the GO tree panel to search using some keywords

Note : In GeneSpring GX GO analysis implementation, all the three component: Molecular Function, Biological Processes and Cellular location are considered together.

On finishing the GO analysis, the *Advanced Workflow* view appears and further analysis can be carried out by the user. At any step in the Guided workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any) and the *Advanced Workflow* view appears.

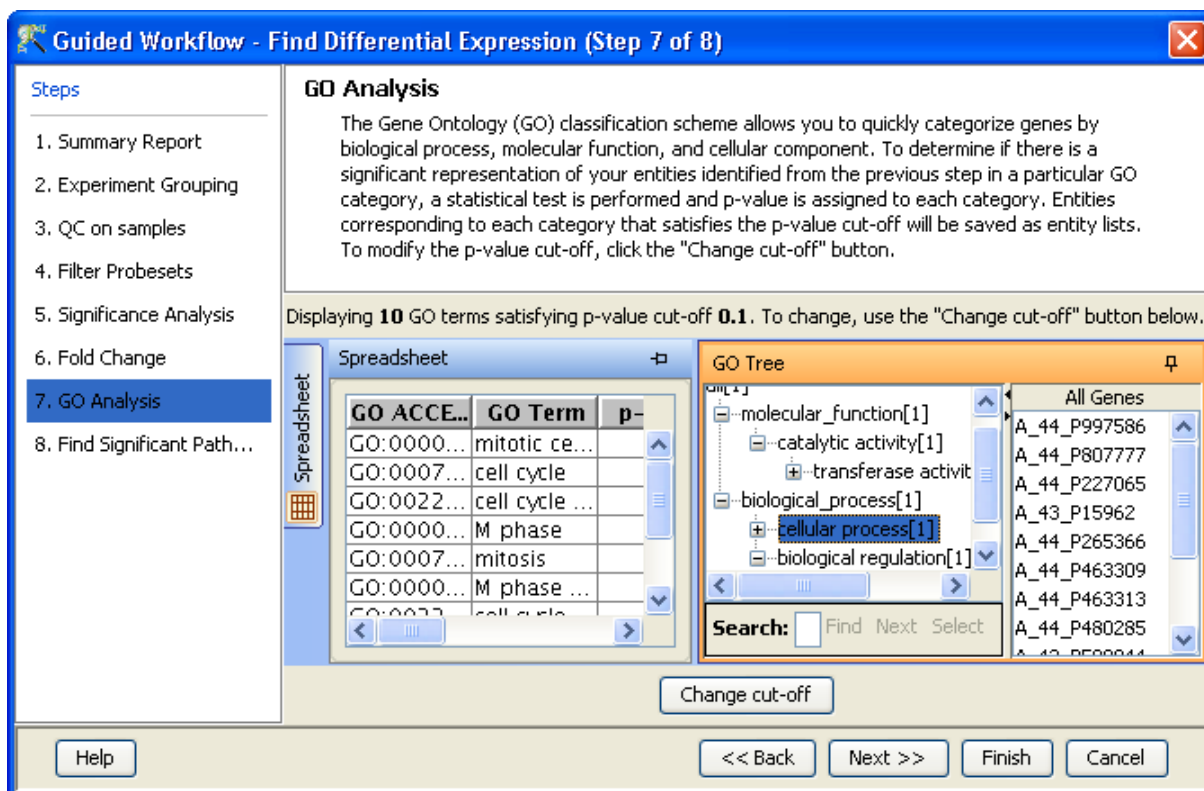


Figure 13.20: GO Analysis

13.3.8 Find Significant Pathways (Step 8 of 8)

This step in the **Guided Workflow** finds relevant pathways from the total number of pathways present in the tool based on similar entities between the pathway and the entity list. The Entity list that is used at this step is the one obtained after the Fold Change (step 6 of 8). This view shows two tables-

- The Significant Pathways table shows the names of the pathways as well as the number of nodes and entities in the pathway and the p-values. It also shows the number of entities that are similar to the pathway and the entity list. The p-values given in this table show the probability of getting that particular pathway by chance when these set of entities are used.
- The Non-significant Pathways table shows the pathways in the tool that do not have a single entity in common with the ones in the given entity list.

The user has an option of defining the p-value cut-off (using *Change cutoff*) and also to save specific pathways using the *Custom Save* option. On clicking, *Finish* the main tool window is shown and further analysis can be carried out by the user. The user can view the entity lists and the pathways created as a result of the Guided Workflow on the left hand side of the window under the experiment in the **Project Navigator**. At any step in the Guided Workflow, on clicking *Finish*, the analysis stops at that step (creating an entity list if any). See figure 13.21.

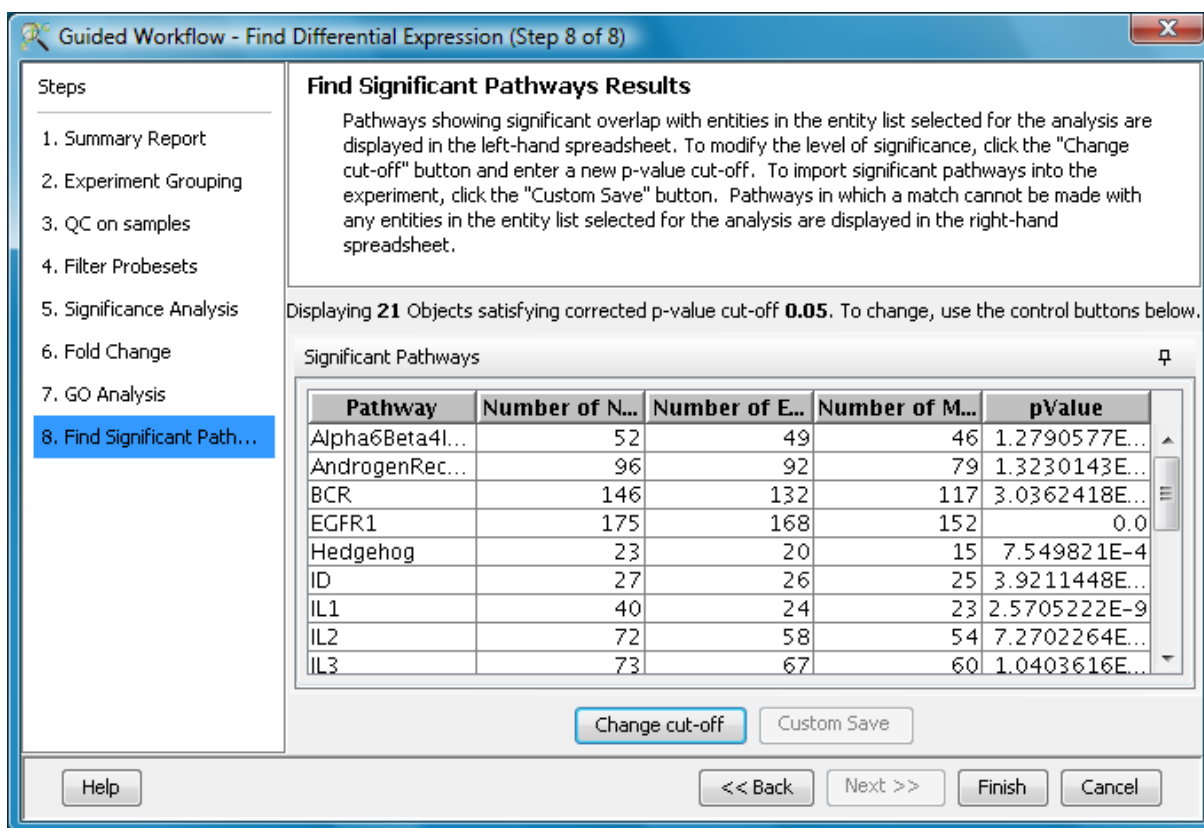


Figure 13.21: Find Significant Pathways

Note: In case the user is using **GeneSpring GX** for the first time, this option will give results using the demo pathways. The user can upload the pathways of his/her choice by using the option **Import BioPAX pathways** under **Tools** in the Menu bar in the main tool window. Later instead of reverting to the Guided Workflow the user can use the option **Find Significant Pathways** in **Results Interpretation** under the same Workflow.

The default parameters used in the *Guided Workflow* are summarized below

13.4 Advanced Workflow

The *Advanced Workflow* offers a variety of choices to the user for the analysis. Flag options can be changed and raw signal thresholding can be altered. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the *Advanced Workflow*, load the data as described earlier. In the **New Experiment Dialog**, choose the **Workflow Type** as Advanced. Clicking on **OK** will open a new experiment wizard which then proceeds as follows:

	Parameters	Parameter values
Expression Data Transformation	Thresholding	1.0
	Normalization	Shift to 75th Percentile
	Baseline Transformation	Not Applicable
	Summarization	Not Applicable
Filter by		
1.Flags	Flags Retained	Present(P)
2.Expression Values	(i) Upper Percentile cutoff	Not Applicable
	(ii) Lower Percentile cutoff	
Significance Analysis	p-value computation	Asymptotic
	Correction	Benjamini-Hochberg
	Test	Depends on Grouping
	p-value cutoff	0.05
Fold change	Fold change cutoff	2.0
GO	p-value cutoff	0.1
Find Significant Pathways	p-value cutoff	0.05

Table 13.8: Table of Default parameters for Guided Workflow

1. New Experiment (Step 1 of 4):

As in case of *Guided Workflow*, either data files can be imported or else pre-created samples can be used.

- For loading new data files, use **Choose Files**.
- If the data files have been previously used in **GeneSpring GX** experiments **Choose Samples** can be used.

Step 1 of 4 of Experiment Creation, the 'Load Data' window, is shown in figure [13.22](#).

2. New Experiment (Step 2 of 4):

Criteria for preprocessing of input data is set here. It allows the user to threshold raw signals to chosen values and select normalization algorithms. The gTotalGeneSignal from FE output which is already background subtracted is brought in. All additional processing steps are performed on this column.

- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value, using which normalization is performed.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the **Available Samples** in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50. The default is set as scale to median of all samples.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The median of the control genes is then used for normalization.

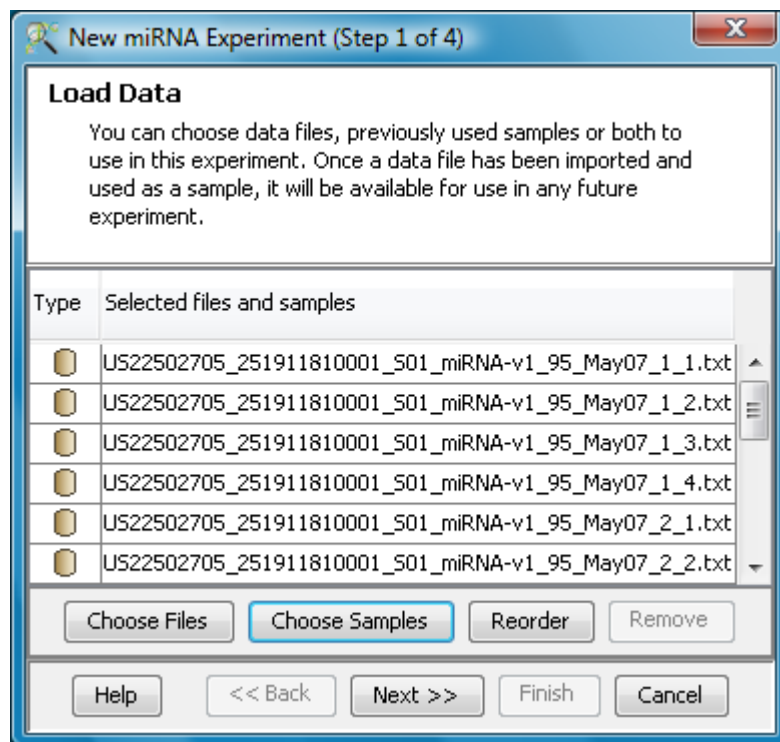


Figure 13.22: Load Data

- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the 'Assign Value' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.
- **Quantile:** On selecting this option, the tool performs Quantile normalization. The user does not have to enter any specifications for this normalization. For details on the above normalization methods, refer to section [Normalization Algorithms](#). Figure 13.23 shows the Step 2 of 4 of experiment creation.

If no normalization is desired, then the option *None* can be chosen.

3. New Experiment (Step 3 of 4):

If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

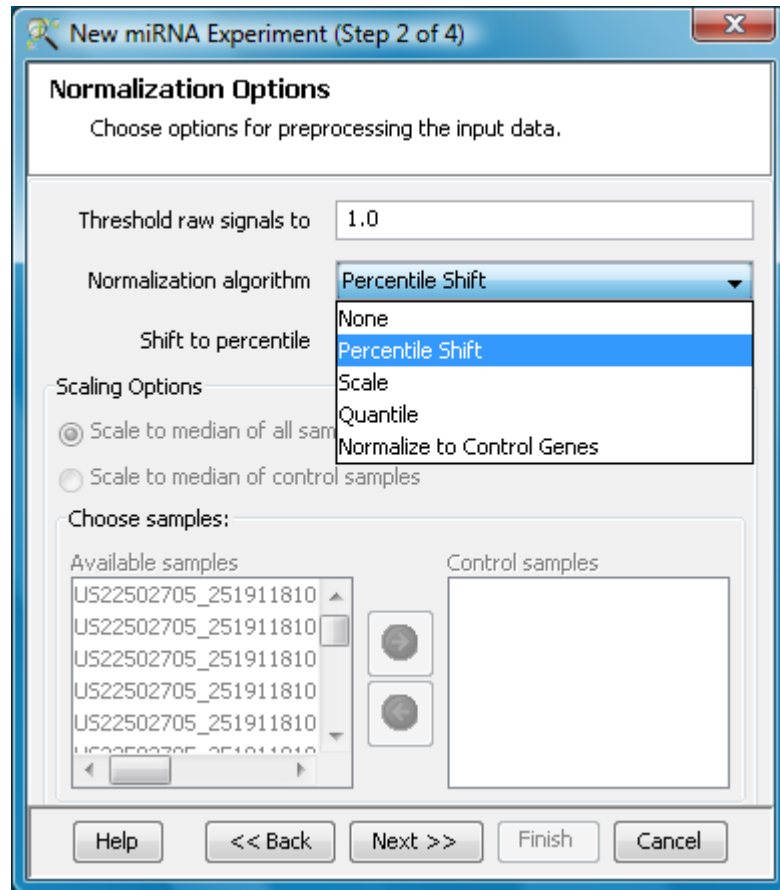


Figure 13.23: Normalization Options

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same. See figure 13.24

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. Such entities will appear under unmatched probe IDs in the experiment notes in the experiment inspector.

4. New Experiment (Step 4 of 4):

Baseline Transformation is carried out row-wise across all samples. This data processing step is particularly useful when visualizing the results in a profile plot or heat map. The baseline transformation options (See figure 13.25), available in **GeneSpring GX** are:

- *Do not perform baseline*
- *Baseline to median of all samples:* For each row (probe), the median of the log summarized values across all the samples is calculated. This value is then subtracted from the probe value for all samples.
- *Baseline to median of control samples:* Here control samples are used to calculate the median value, for each probe. This value is then subtracted from the probe value for all samples. The controls could be an individual control for each sample or it could be a set of controls. Alternatively, a set of samples can be used as controls for all samples. For specifying the control

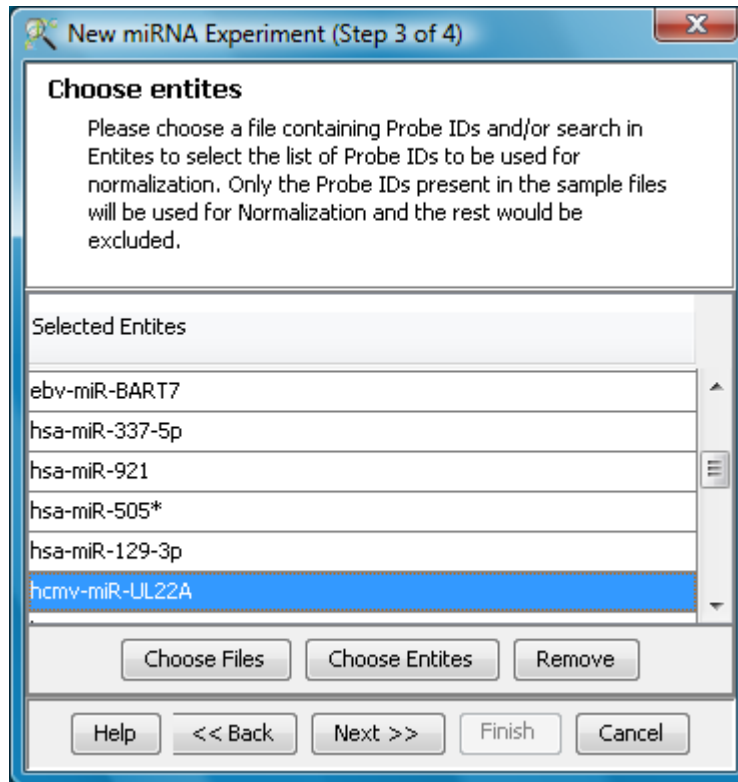


Figure 13.24: Choose entities

for a sample, select the sample and click on **Assign value**. This opens up the **Choose Control Samples** window from where the samples designated as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. See figure 13.26. Click on **Ok**. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Clicking **Finish** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to **View** in Toolbar.

Once an experiment is created, the *Advanced Workflow* steps appear on the right hand side. Following is an explanation of the various workflow links:

13.4.1 Experiment Setup

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the Advanced Workflow,

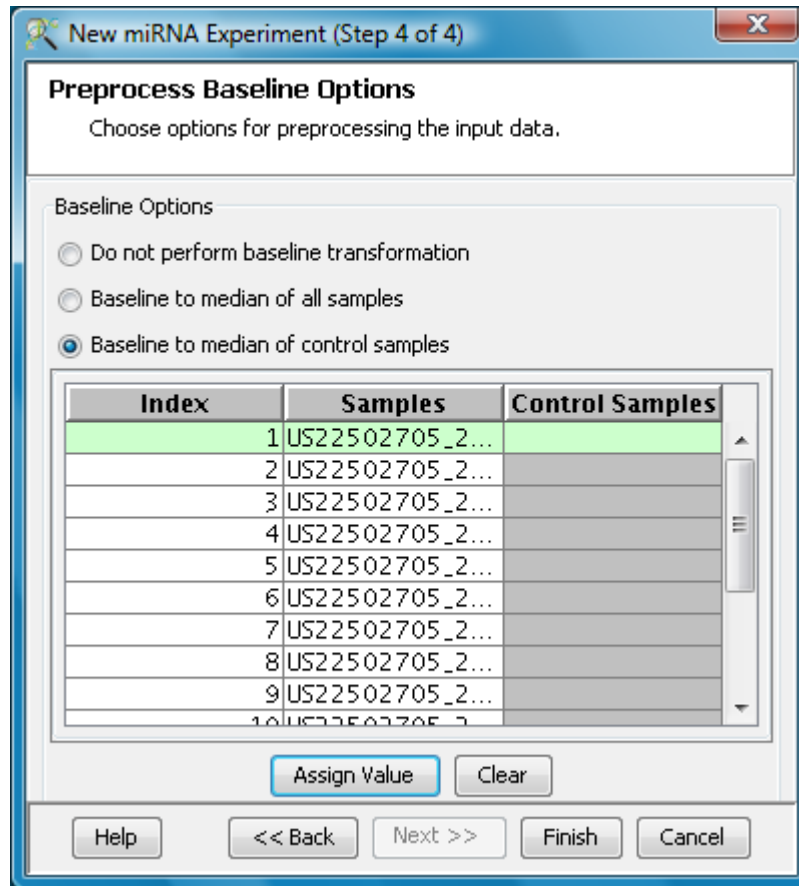


Figure 13.25: Baseline Transformation

the method of analysis, the details of the algorithms used and the interpretation of results

- **Experiment Grouping:** *Experiment Parameters* defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)

13.4.2 Quality Control

- **Quality Control on Samples:**

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows three tiled windows:

- 3D PCA Scores, Correlation Plots and Correlation Coefficients.

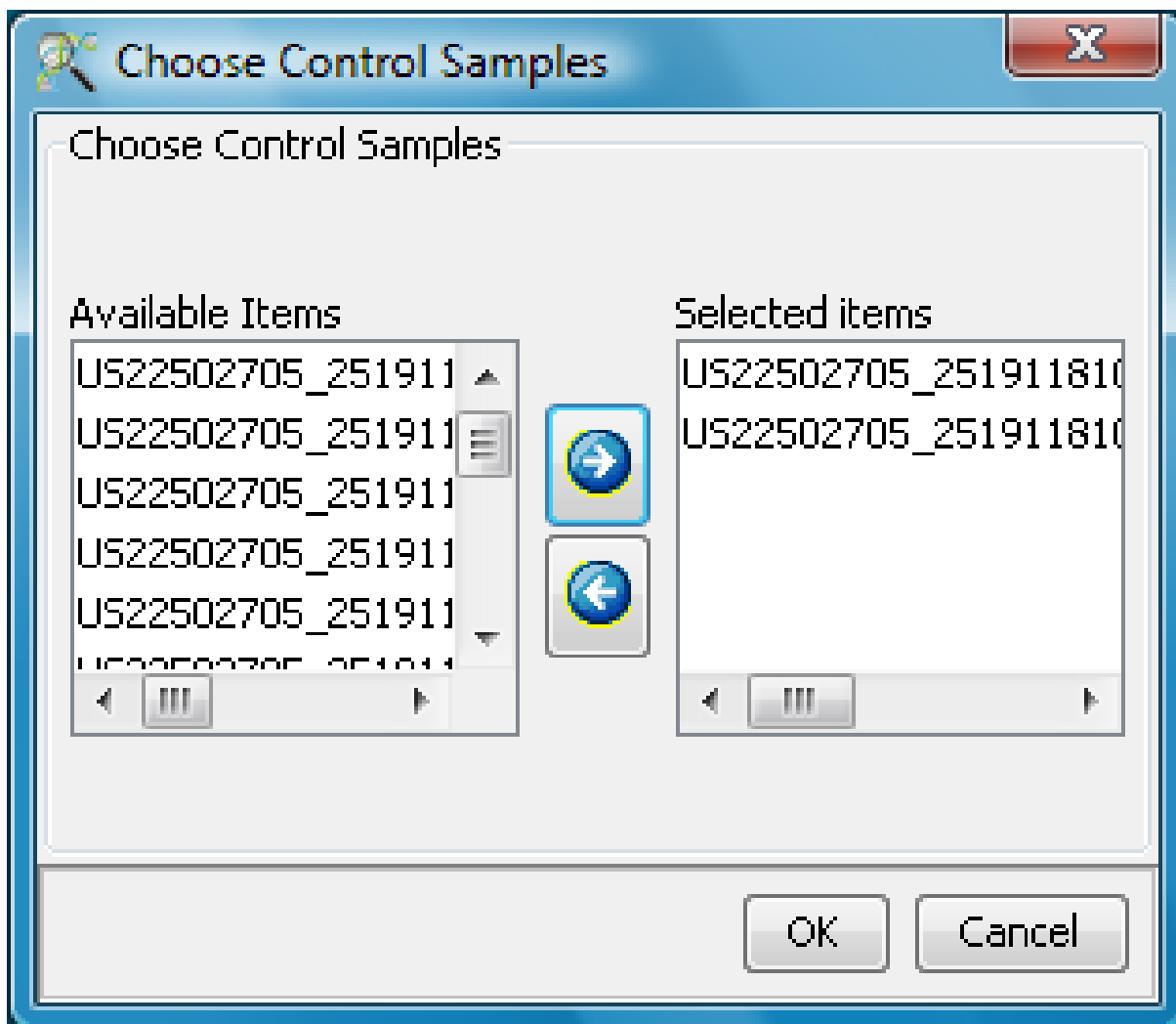


Figure 13.26: Selection of Controls

- Quality Metrics Report, Quality Metrics plot and Experiment Grouping tabs.
- Legend.

Figure 13.27 has the 3 tiled windows which reflect the QC on samples.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click** → **Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The **Correlation Plots** shows the correlation analysis across arrays. It finds the correlation coef-

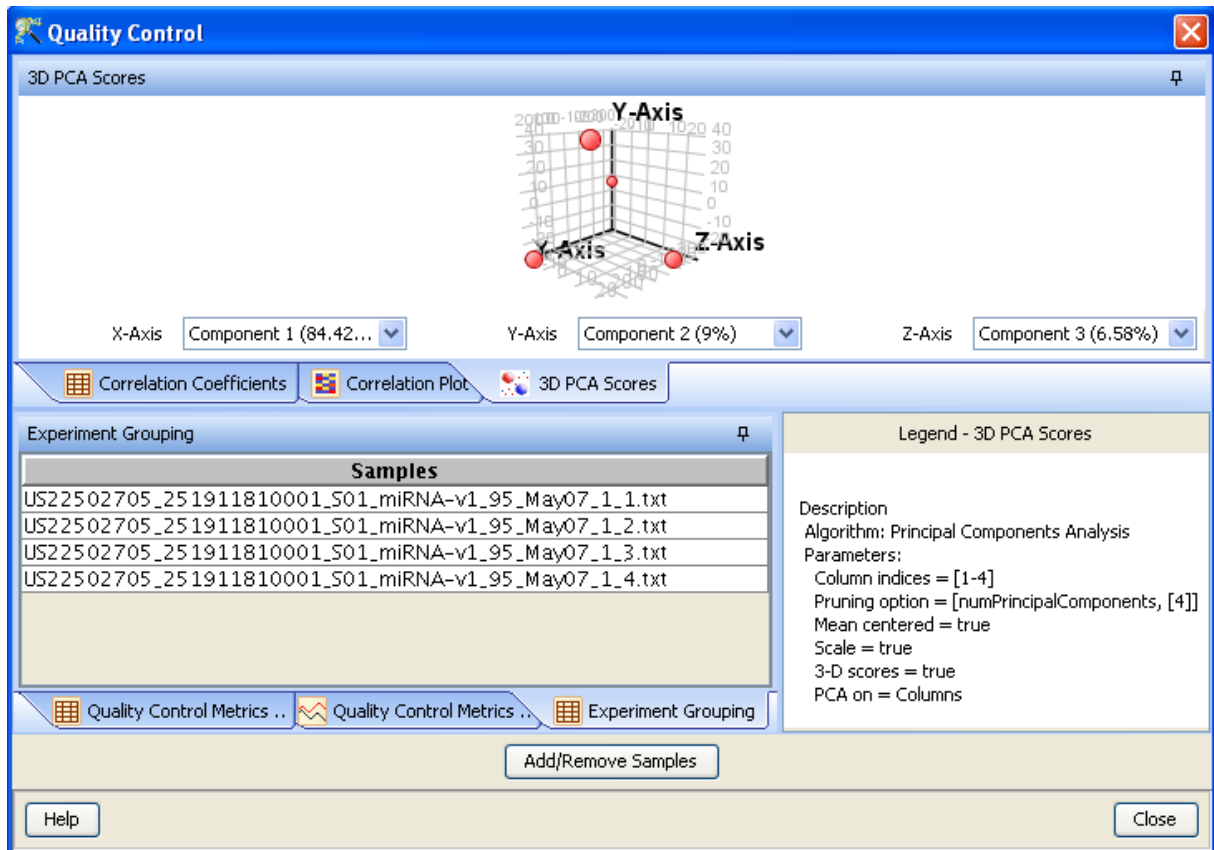


Figure 13.27: Quality Control

efficient for each pair of arrays and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient. A value of 1.0 indicates maximum correlation.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$

The heatmap is colorable by *Experiment Factor* information via Right-Click → Properties. Similarly, the intensity levels in the heatmap are also customizable.

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

The metrics report helps the user evaluate the reproducibility and reliability of the microarray data. The quality metrics scores are obtained directly from the sample file. A brief description is given below:

- Additive error (AddErrorEstimateGreen): measures on feature background noise. Should be <5, 5~12 is concerning, >12 is bad

- % Feature Population Outlier (AnyColorPrentFeatPopnOL): Measures % of features that are called population outliers (and therefore excluded from analysis) Should be less than 8%, >~15% is bad
- NonControl %CV of BGsubtracted Signal (gNonCtrlMedPrentCVBGSubSig): Measures uniformity of signals across feature replicates Should be <10%, >~15% is bad, -1 is bad
- 75% ile Total Gene Signal (gTotalSignal75pctile): Measures overall intensity of non control probes. This metric is HIGHLY sample dependant, but should be consistent for well behaving samples of similar type.

More details on this can be obtained from the Agilent Feature Extraction Software(v9.5) Reference Guide, available from <http://chem.agilent.com>.

Quality controls Metrics Plot shows the QC metrics present in the QC report in the form of a plot.

Experiment Grouping shows the parameters and parameter values for each sample.

The third window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using *Add/Remove Samples* button. Once a few samples are removed, re-normalization and baseline transformation(if chosen) of the remaining samples is carried out again. The samples removed earlier can also be added back. Click on *OK* to proceed.

- **Filter Probe Set by Expression:**

Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)

- **Filter Probe Set by Flags:**

In this step, the entities are filtered based on their flag values as either P(present) or A(absent). Information pertaining to the flags is present in the data file. **GeneSpring GX** considers the "gIs-GeneDetected" as the flag column and marks entities having '0' as *Absent* and '1' as *Present*. This process is done in 4 steps:

1. Step 1 of 4 : *Entity list and Interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window. This is seen in figure [13.28](#)
2. Step 2 of 4: This step is used to set the filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in **Retain Entities** box. See figure [13.29](#).
3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. See figure [13.30](#).
4. Step 4 of 4: Click **Next** to save the entity list created as a result of this analysis. See figure [13.31](#).

13.4.3 Analysis

- **Statistical Analysis**

For details refer to section [Statistical Analysis](#) in the advanced workflow.

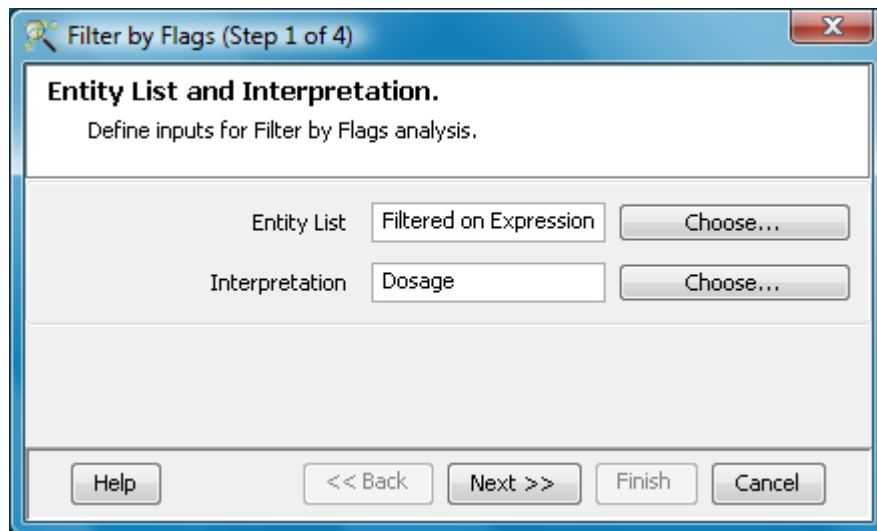


Figure 13.28: Entity list and Interpretation

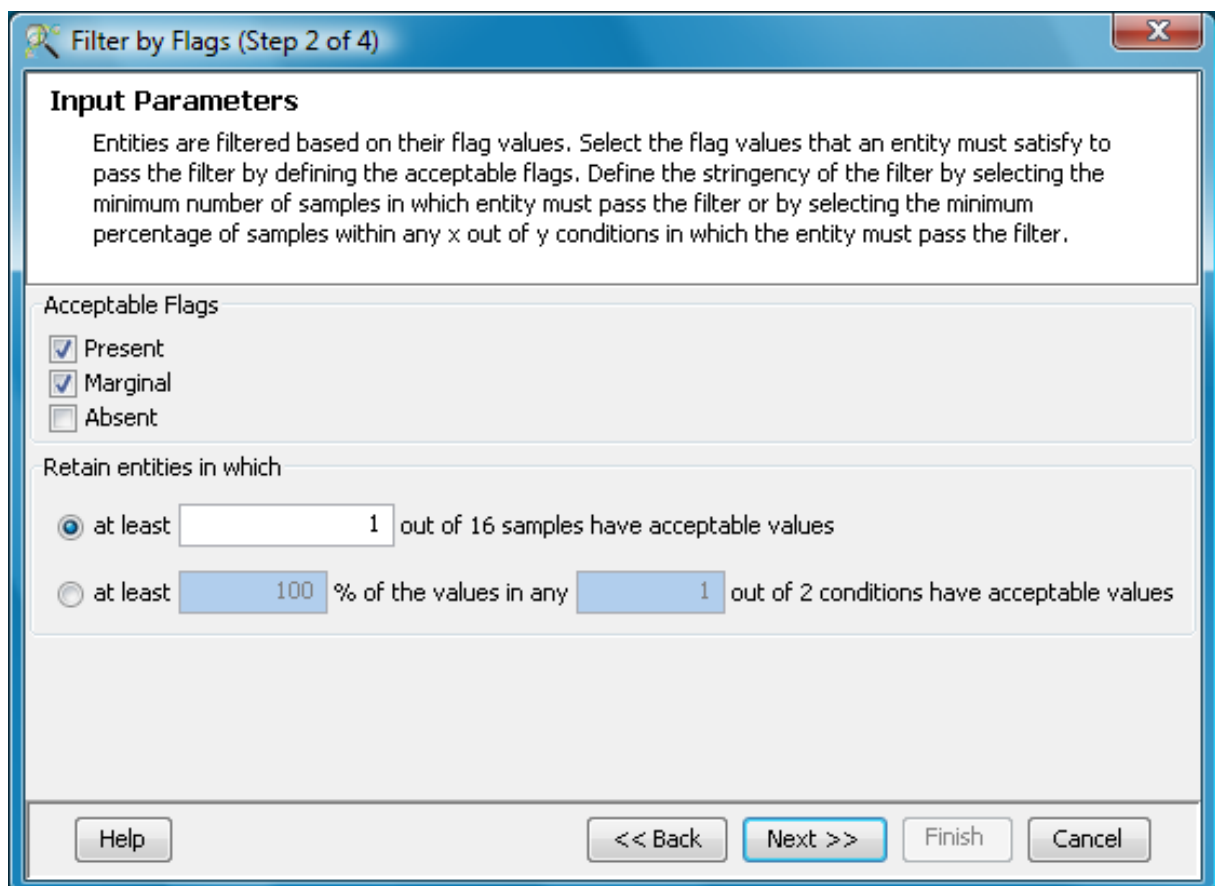


Figure 13.29: Input Parameters

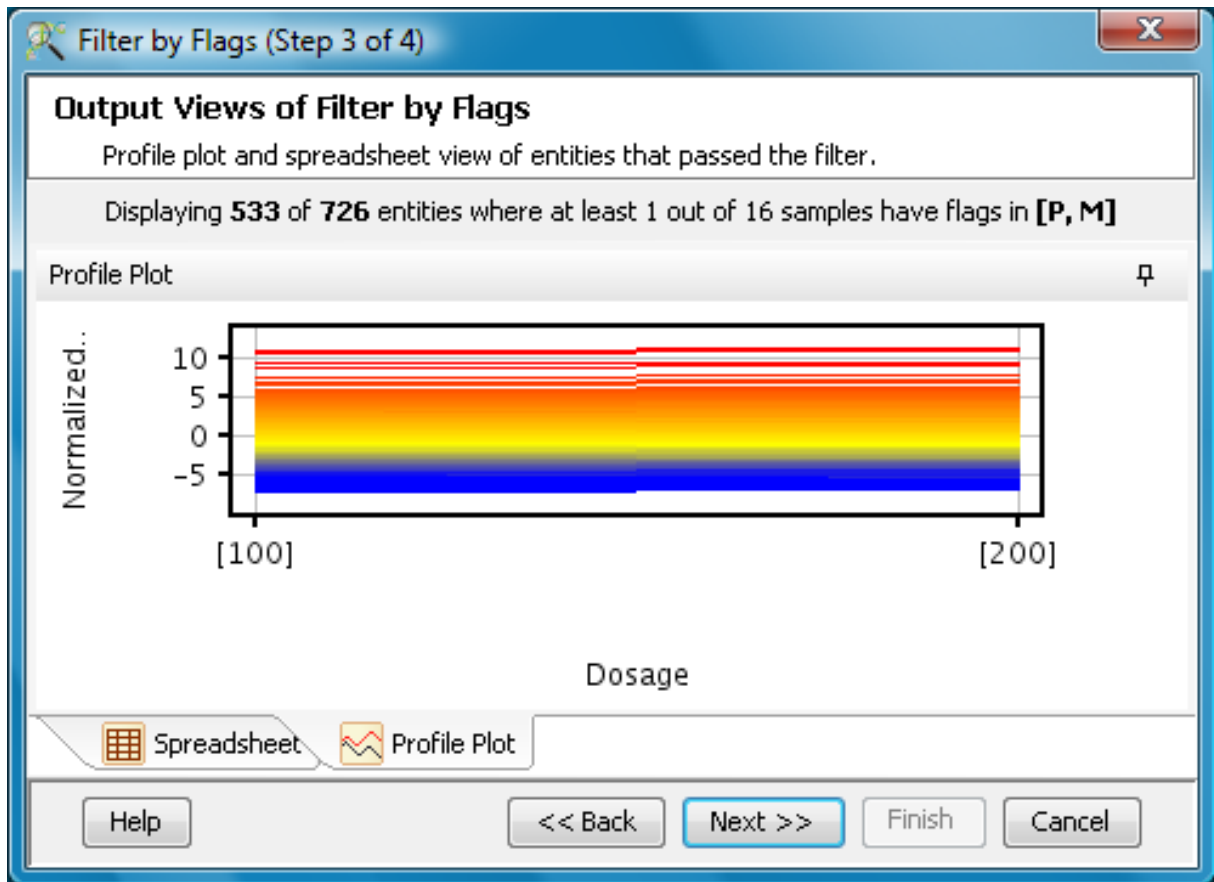


Figure 13.30: Output Views of Filter by Flags

- **Filter on Volcano Plot**
For details refer to section [Filter on Volcano Plot](#)
- **Fold Change**
For details refer to section [Fold Change](#)
- **Clustering**
For details refer to section [Clustering](#)
- **Find Similar Entities**
For details refer to section [Find Similar Entities](#)
- **Filter on Parameters**
For details refer to section [Filter on Parameters](#)
- **Principal Component Analysis**
For details refer to section [PCA](#)

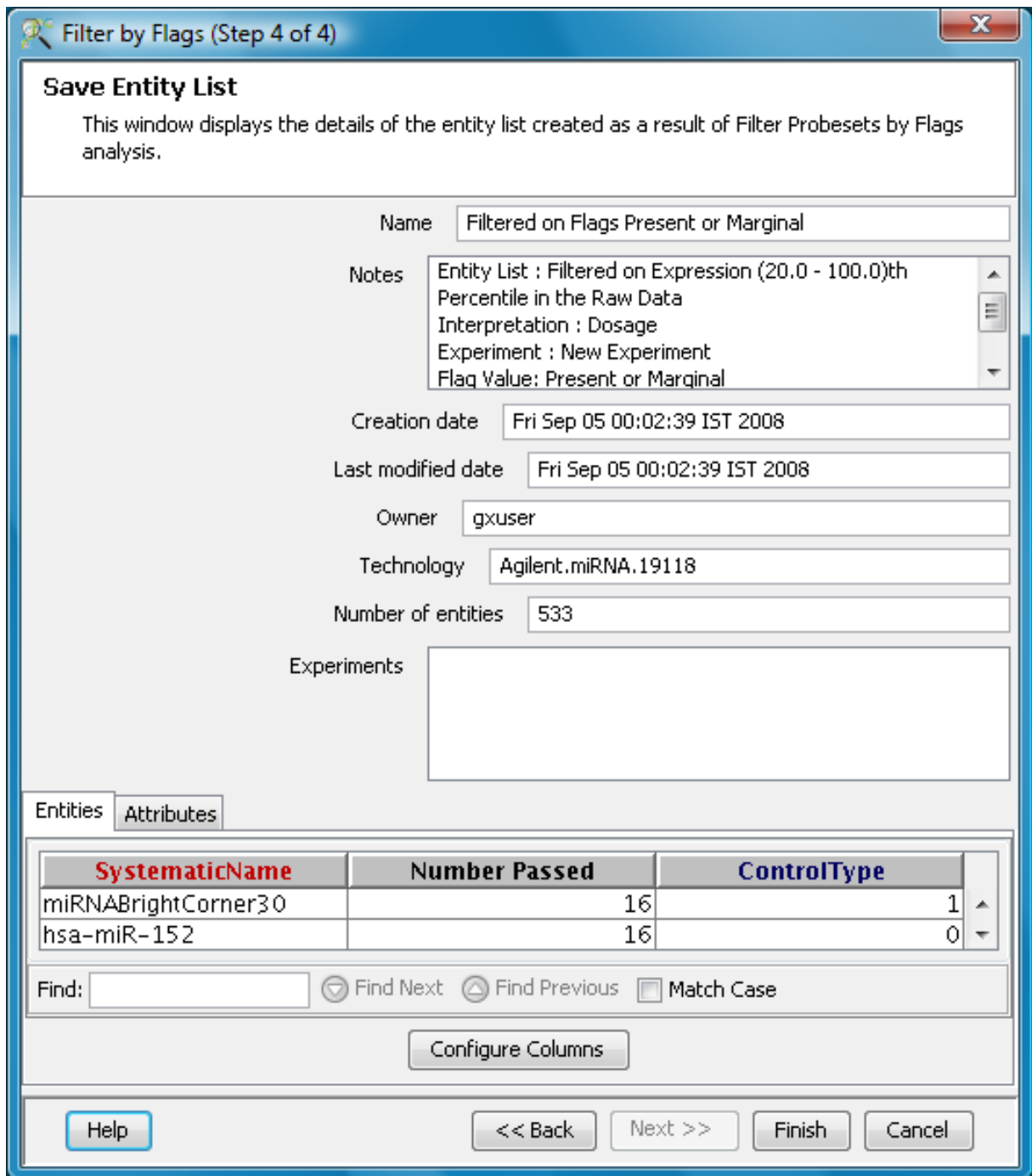


Figure 13.31: Save Entity List

13.4.4 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

13.4.5 Results

- **Gene Ontology (GO) analysis**
GO is discussed in a separate chapter called [Gene Ontology Analysis](#).
- **Gene Set Enrichment Analysis (GSEA)**
Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).
- **Gene Set Analysis (GSA)**
Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).
- **Pathway Analysis**
Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).
- **Find Similar Entity Lists**
This feature is discussed in a separate section called [Find Similar Entity Lists](#)
- **Find Significant Pathways**
This feature is discussed in a separate section called [Find Significant Pathways](#).
- **Launch IPA**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Import IPA Entity List**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Extract Interactions via NLP**
This feature is discussed in detail in the chapter [Pathway Analysis](#).

13.4.6 TargetScan

GeneSpring GX miRNA workflow not only identifies significant miRNAs, but also facilitates identification of the target genes regulated by the miRNAs. This is possible due to the TargetScan database <http://www.targetscan.org/> which is integrated in **GeneSpring GX** .

TargetScan allows identification of mRNA targets for any specific miRNA, based on the context percentile and the database which are user defined. Context percentile is derived from context score and has been described as:

”Sites within 15 nt of a stop codon are flagged because these are typically not effective. The context of each of the remaining sites has been evaluated and scored considering the following four features:

- site-type contribution: reflects the type of seed match (8mer, 7mer-m8, and 7mer-1A)
- 3' pairing contribution: reflects consequential miRNA-target complementarity outside the seed region
- local AU contribution: reflects transcript AU content 30 nt upstream and downstream of predicted site
- position contribution: reflects distance to nearest end of annotated UTR of target

With all four features, a more negative score is associated with a more favorable site. The context score is the sum of the above scores, and the context score percentile is the percentile rank of each site compared to all sites for this miRNA family. Thus a high context score percentile (between 50 and 100) shows that a specific site is more favorable than most other sites of this miRNA.” (Taken from- <http://www.TargetScan.org/docs/help.html>)

The other criteria which determines target selection is the database. There are two databases, conserved and non-conserved. For miRNA target sites, conservation is defined using the conserved branch length which is based on the sum of phylogenetic branch lengths between species that contain a site and also dependent on site type and UTR conservation.

The conserved branch length score (Friedman et al., 2008) is the sum of phylogenetic branch lengths between species that contain a site. To help control for individual UTR conservation, 3' UTRs were separated by conservation rate into ten equally sized bins, and a unique set of branch lengths based on 3' UTR sequence alignments was constructed for each bin. Site conservation is defined by conserved branch length, with each site type having a different threshold for conservation:

- 8mer: 0.8
- 7mer-m8: 1.3
- 7mer-1A: 1.6

(Taken from - http://www.targetscan.org/cgi-bin/targetscan/vert_50/view_gene.cgi?taxid=9606&gs=TNKS2&members=miR-1/206)

Note that in TargetScan, definitions of conservation can apply to (1) miRNA families and (2) miRNA target sites. In the context of **GeneSpring GX**, it is only the miRNA target sites and their relevant databases described above are relevant.



Figure 13.32: Workflow Navigator-TargetScan

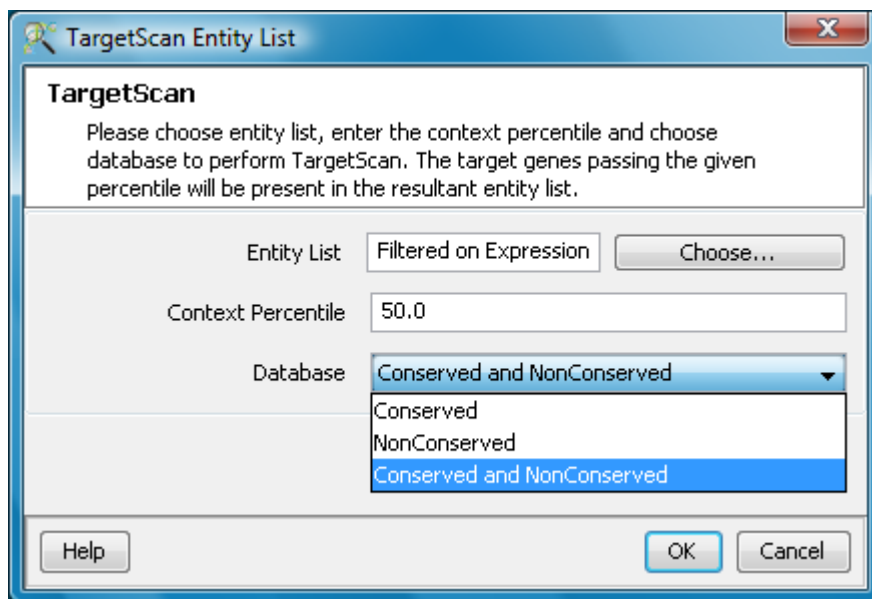


Figure 13.33: Inputs for TargetScan

TargetScan can be accessed from *Workflow Navigator* → *Results Interpretation* → *TargetScan*. See figure 13.32. The tool downloads the organism's TargetScan database when this option is used for the first time. Here a number of user defined inputs are needed. See figure 13.33

- **Entity List:** This would be the list of miRNAs whose targets are needed.
- **Context Percentile:** Default is set at 50.
- **Database:** Either of the 3 options: Conserved, Non-Conserved or both can be chosen

TargetScan creates a list of the targets for the entities under the original entity list. Analysis methods such as **GO Analysis** accept only target entity lists as input. Since the miRNA technology is not associated

with the target gene annotations, it is necessary to have the appropriate organism's **Biological Genome** created. Using this feature, the GO terms are obtained for the TargetScan list. Similarly, for Pathway Analysis, the Entrez IDs are obtained from **Biological Genome**. While it is possible to perform Pathway Analysis, clicking on the TargetScan entity list will not highlight the TargetScan entities present on a pathway. For more information on creating a genome refer to the section on [Biological Genome](#)

13.4.7 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)
- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)
- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)
- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

Chapter 14

Analyzing Real Time PCR Data

Real Time PCR (RT-PCR) also called Quantitative PCR (qPCR) is used to rapidly measure the quantity of DNA, cDNA, or RNA present in a sample. It is the most sensitive technique for mRNA detection and quantization currently available. Compared to the two other commonly used techniques for quantifying mRNA levels, Northern blot analysis and RNase protection assay, RT-PCR can be used to quantify mRNA levels from much smaller samples. In fact, this technique is sensitive enough to enable quantitation of RNA from a single cell.

GeneSpring GX supports all version of the ABI's 7900HT RT-PCR system. The columns that are imported into **GeneSpring GX** from the original data file are the Sample, Detector, Task and Ct. In addition, the tool also creates a Gene symbol and a synonyms(of the GeneSymbol) column

14.1 Running the Real Time PCR Workflow

Upon launching **GeneSpring GX** , the startup is displayed with 3 options.

1. Create new project
2. Open existing project
3. Open recent project

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting *Create new project*, a window appears in which details (Name of the project and Notes) can be recorded. Press **OK** to proceed. An Experiment Selection Dialog window then appears with two options:

1. Create new experiment
2. Open existing experiment

Selecting *Create new experiment* allows the user to create a new experiment (steps described below). *Open existing experiment* allows the user to use existing experiments from any previous projects in the current project. Choosing *Create new experiment* opens up a **New Experiment dialog** in which experiment name can be assigned. The experiment type should then be specified. The drop-down menu gives the user the option to choose between the Affymetrix Expression, Affymetrix Exon Expression, Illumina Single Color, Agilent One Color, Agilent Two Color, Real Time PCR, Pathway, Generic Single Color and Two Color experiment types.

Upon clicking **OK**, the Real Time PCR experiment creation wizard appears. This wizard requires details such as name of the technology, organism under study and the sample files for experiment creation. See figure 14.1

The next step allows the user to perform baseline transformation. See figure 14.2. Baseline Transformation is carried out row-wise across all samples. This data processing step is particularly useful when visualizing the results in a profile plot or heat map. The baseline transformation options, available in **GeneSpring GX** are:

- *Do not perform baseline*
- *Baseline to median of all samples:* For each row (probe), the median of the log summarized values across all the samples is calculated. This value is then subtracted from the probe value for all samples.
- *Baseline to median of control samples:* Here control samples are used to calculate the median value, for each probe. This value is then subtracted from the probe value for all samples. The controls could be an individual control for each sample or it could be a set of controls. Alternatively, a set of samples can be used as controls for all samples. For specifying the control for a sample, select the sample and click on *Assign value*. This opens up the **Choose Control Samples** window from where the samples designated as *Controls* should be moved from the *Available Items* box to the *Selected Items* box. Click on **Ok**. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Clicking **Finish** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to **View** in Toolbar.

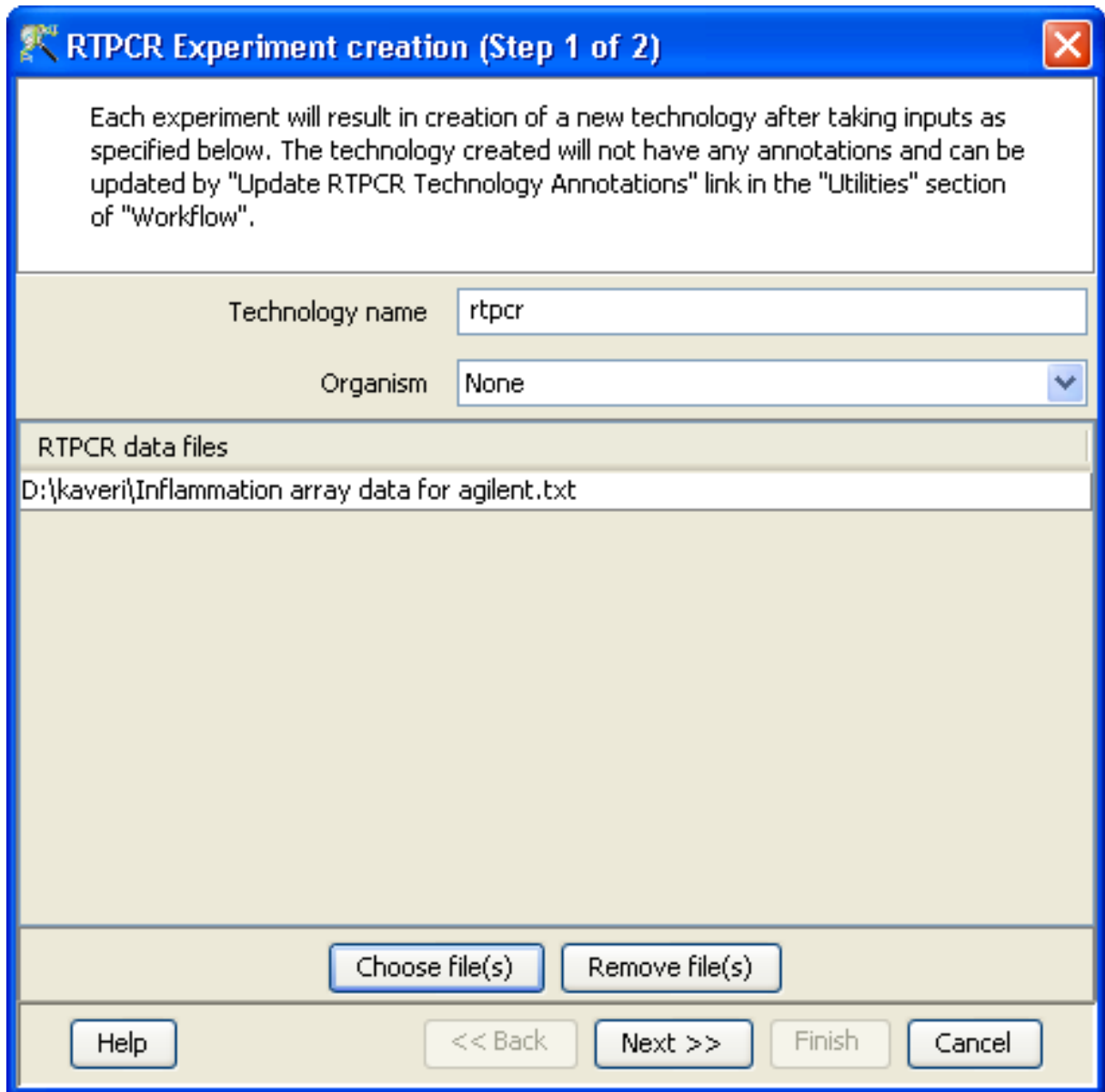


Figure 14.1: Experiment Creation

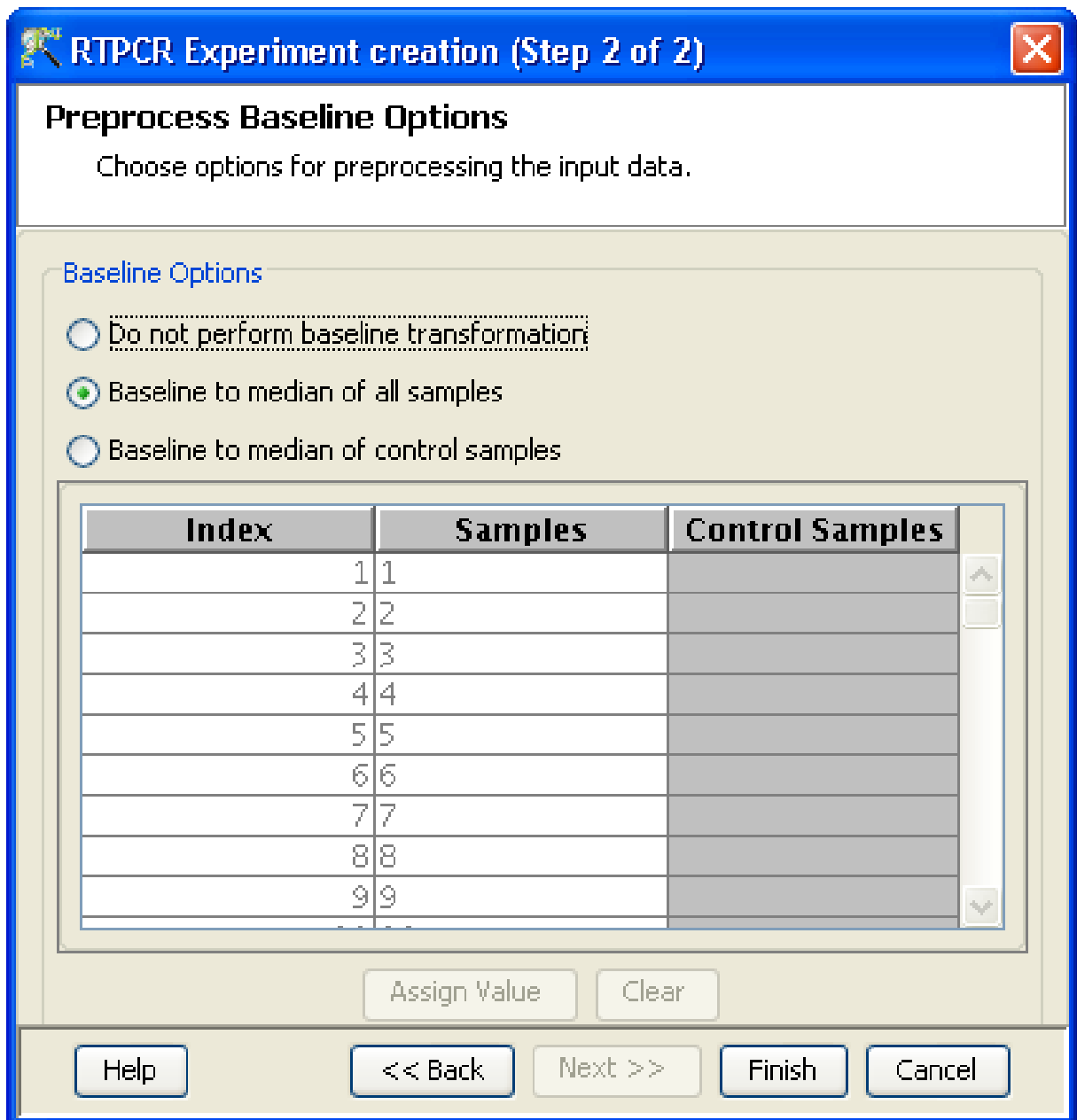


Figure 14.2: Baseline Transformation Options

14.1.1 Technology Creation in RT-PCR experiments

For each experiment, a new Technology is created within **GeneSpring GX**. The Technology name is in the format ABL.RTPCR.RQ(version)-(name) where version corresponds to the supported version of the RQ Manager software and name corresponds to what is provided by the user during the experiment creation process. **GeneSpring GX** can detect the following Annotations - Detector, Assay ID, Gene Symbol and Synonyms for the Technology out of the supported RQ data files based on the following guidelines :

All the above mentioned Annotations are derived from the Detector column and the following format is assumed

Synonym(Gene Symbol)-Assay ID: Please note that only the Assay ID is mandatory in the above format.

Here are a few examples

- If the user wants to bring in the Gene Symbol GS1 with an Assay ID Assay1 then the Detector column should be specified as: (GS1)-Assay1
- If the user wants to bring in the Synonym S1, Gene Symbol GS1 with an Assay ID Assay1 then the Detector column should be specified as: S1(GS1)-Assay1
- For example, if the value in the Detector column is ATIR(AGTRI)-HS00241341_m1, then AGTRI will be the Gene Symbol, ATIR will be the synonym and HS00241341 will be the Assay ID

14.1.2 Data Processing

1. **File formats:** The files should be in text (.txt) format.
2. **Raw:** The term "raw" signal values refer to the data after averaging Avg Ct or Ct Avg column within a sample (summarization).
3. **Normalized:** The term Normalized signal value refers to a difference between the summarized Averaged counts of Endogenous controls and the target within a sample. It also reflects the baseline transformation performed.
4. **Treatment of on-chip replicates:** Replicates of a target are averaged to compute their total intensity values as described above.
5. **Flag values:** Not applicable.
6. **Treatment of Control probes:** The control probes that are taken into account are the endogenous control probes which are identified by the "Task" column in the original data file.
7. **Empty Cells:** Empty cells might be present in the intensity values column for certain genes in the data file. These genes are brought in **GeneSpring GX**. These can be removed from the entity lists during analysis from *Utilities* → *Remove Entities with missing signal values*.

8. **Sequence of events:** The sequence of events involved in the processing of the data files is: summarization, normalization and baseline transformation.

14.1.3 Experiment Setup

Once an experiment is created, the **Advanced Workflow** steps appear on the right hand side. Following is an explanation of the various workflow links

- **Quick Start Guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading files into **GeneSpring GX**, the **Advanced Workflow**, the method of analysis, the details of the algorithms used and the interpretation of results
- **Experiment Grouping: Experiment Parameters** define the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)

14.1.4 Quality Control

- **Quality Control on Samples:**

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, samples can be removed from the analysis. The QC view shows four tiled windows:

- Correlation plots and Correlation coefficients
- PCA scores
- Experiment grouping
- Legend

Figure 14.3 has the 4 tiled windows which reflect the QC on samples.

The *Correlation Plots* shows the correlation analysis across samples. It finds the correlation coefficient for each pair of samples and then displays these in textual form as a correlation table as well as in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**.
Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$

The heatmap is colorable by Experiment Factor information via Right-Click → Properties. Similarly, the intensity levels in the heatmap are also customizable.

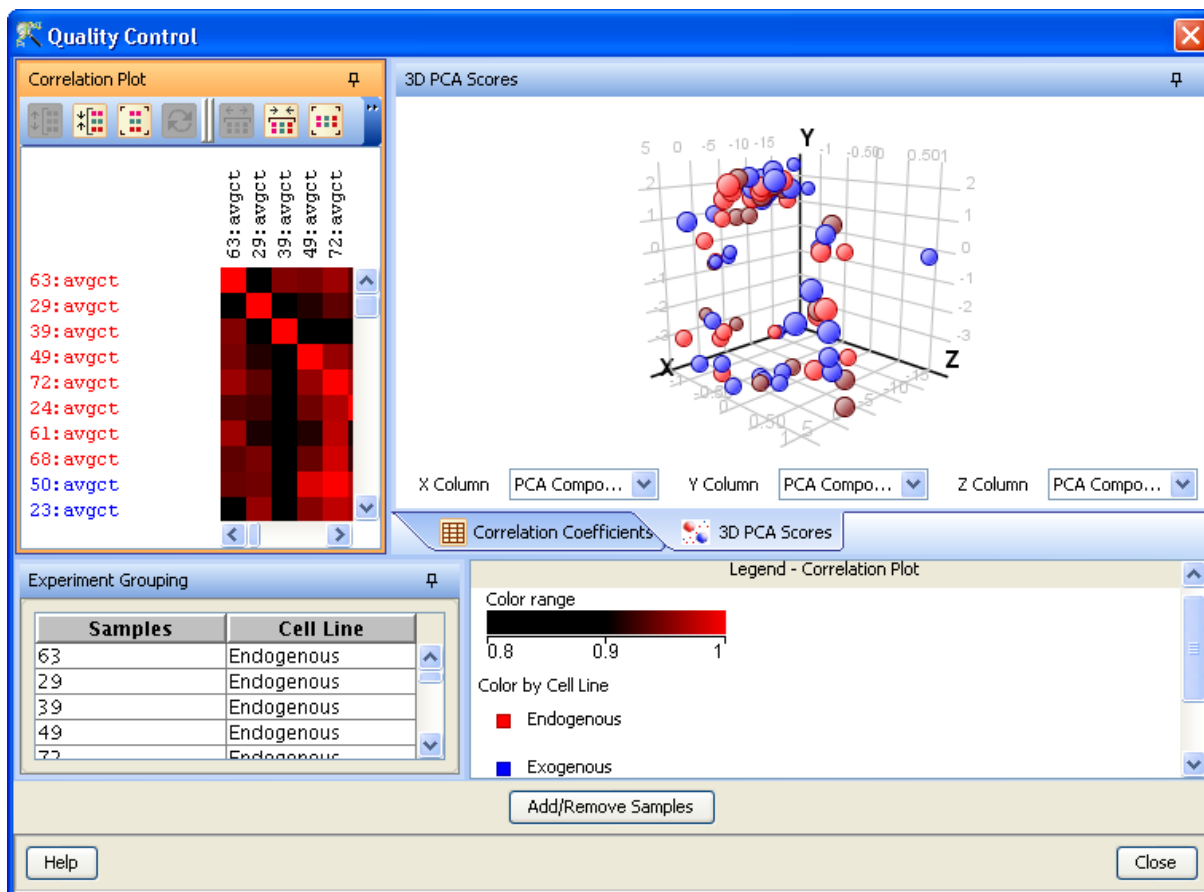


Figure 14.3: Quality Control

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

Experiment Grouping shows the parameters and parameter values for each sample.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according in the decreasing order of significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The fourth window shows the legend of the active QC tab.

Unsatisfactory samples or those that have not passed the QC criteria can be removed from further analysis, at this stage, using **Add/Remove Samples** button. Once samples are removed,

re-normalization and baseline transformation of the remaining samples are carried out again. The samples removed earlier can also be added back. Click on **OK** to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags:** This is not applicable as flags are not created in this experiment type.
- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on [Filter Probesets on Data Files](#)
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

14.1.5 Analysis

- **Statistical Analysis**

For details refer to section [Statistical Analysis](#) in the advanced workflow.

- **Filter on Volcano Plot**

For details refer to section [Filter on Volcano Plot](#)

- **Fold Change**

For details refer to section [Fold Change](#)

- **Clustering**

For details refer to section [Clustering](#)

- **Find Similar Entities**

For details refer to section [Find Similar Entities](#)

- **Filter on Parameters**

For details refer to section [Filter on Parameters](#)

- **Principal Component Analysis**

For details refer to section [PCA](#)

14.1.6 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

14.1.7 Results

- **Gene Ontology (GO) analysis**

GO is discussed in a separate chapter called [Gene Ontology Analysis](#).

- **Gene Set Enrichment Analysis (GSEA)**

Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).

- **Gene Set Analysis (GSA)**

Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).

- **Pathway Analysis**

Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).

- **Find Similar Entity Lists**

This feature is discussed in a separate section called [Find Similar Entity Lists](#)

- **Find Significant Pathways**

This feature is discussed in a separate section called [Find Significant Pathways](#).

- **Launch IPA**

This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Import IPA Entity List**


This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).

- **Extract Interactions via NLP**

This feature is discussed in detail in the chapter [Pathway Analysis](#).

14.1.8 Utilities

Import Entity List from file

This option allows the user to bring any entity list of interest into **GeneSpring GX**. Typically the entity list is a list of probeset IDs, gene symbols, entrez ids etc along with associated data, all specified in a file in .txt, .csv, .xls, or .tsv formats. Once imported, this list will be added as a child to the 'Imported Lists' folder in the Experiment Navigator. The Entity List could be in the form of gene symbols or Probe set IDs or any other id type present in the technology of the active experiment. The *Import Entity List* dialog can be started either from the Utilities section of the workflow or by clicking on the Import Entity List from File  icon on the toolbar. The dialog consists of four fields:

Choose File - This asks the user to specify the path of the file to be imported.

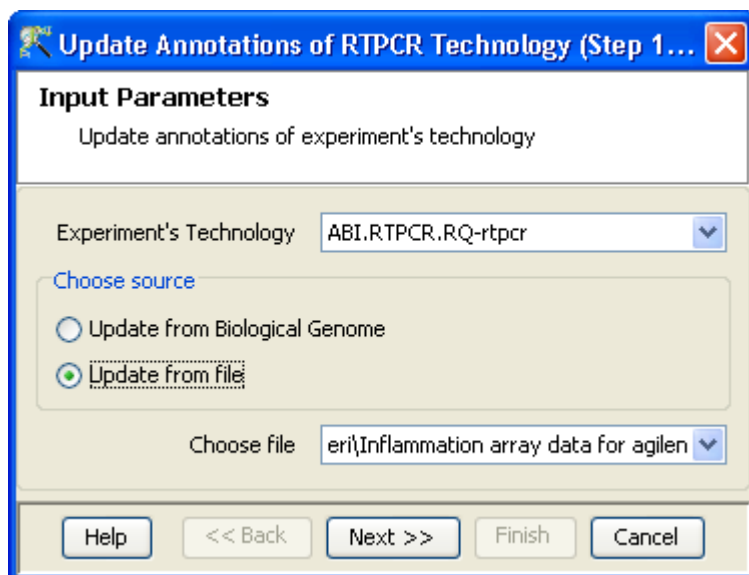


Figure 14.4: Input Parameters

Choose file column to match - Here the user has to choose a column that is present in the imported file. This column will be used to merge the file with the entities in the experiment.

Choose technology column to match - The column to be imported can be either the probeset ID, the UniGene Id or any other id type present in the technology for this experiment. Choose the appropriate mark from the drop-down menu.

Choose columns to import - any other data columns to be imported from the input file can be specified here. These additional columns can be brought in only if the column chosen for matching above is the Probeset ID (or alternatively, whatever is the ID column in the technology for this experiment).

Update RT-PCR Technology Annotations

This option enables the user to update the annotations of the created technology. Selecting this option, opens up a wizard having two steps:

1. **Step 1 of 2** - Here the user specifies the technology as well as the source from which it has to be updated. See figure 14.4. The technology can be updated either from a file or from the **Biological Genome** of that organism. If the **Biological Genome** of that organism does not exist, then the user can create a genome from *Annotations* → *Create Biological Genome*. For more details on the creation of a genome, refer to [Biological Genome](#). If the user chooses to update from a file, then it should be chosen accordingly via the *Choose file* option. The file from which the update is to be performed has to be in a tabular format.
2. **Step 2 of 2** - The annotation columns are merged with the existing technology using a technology identifier. This step asks the user to specify the identifier and to choose the column to be updated

from the annotation file/genome. While specifying the columns, column marks should be assigned. See figure 14.5. It is recommended that the user chooses a column with unique values (for e.g : Entrez-ID) as the identifier. Three kinds of updates are possible:

- Append to the existing information,
- Overwrite
- Fill in the missing values.

Appending the values will retain the original value as well as add the new value. Overwrite will replace the original value with the newer one, whereas fill in missing values will add values at places where previously there were none.

Remove Entities with missing signal values

This option allows the user to remove entities which have missing values in the data file. This usually occurs in the case of custom files. This is important as Clustering and Class Prediction analysis require entity lists with 'no' missing values.

Filter on Entity List

This utility allows user to filter an Entity list using its annotations and list associated values. The filter can be set by defining a search field, a search condition like equals or starts with, and a value for the search field, as applicable. Multiple searches can be combined using OR or AND condition. Filter on Entity List opens a four step wizard.

The *Filter on Entity List* dialog can be started from the Utilities section of the workflow.

Step 1 of 4 : Allows selection of entity list

Step 2 of 4 : Allows defining the filter conditions using three fields Search field, condition and search value. Search field shows all the annotations and list associated values as drop down; depending on the search field, the condition can be a string like equals, does not equal, starts with, ends with, includes or their numerical equivalents; the search value will allow the desired value (either string or a number, depending on the search field) to be input.. More search conditions can be added/removed using the Add/Remove button. There is also a functionality to combine different search conditions using OR or AND conditions.

Step 3 of 4 : The filter results are displayed as a table in this step. Those entities that satisfy the filter conditions are selected by default. All the entities will be selected if the filter conditions are not valid. The selections in the result page can be modified by ctrl-click.

Step 4 of 4 : Allows saving the filtered entity list. Here, the columns in the entity list can be configured before saving. Finish will import the filtered entity lists as a child node under the original entity list in the experiment.

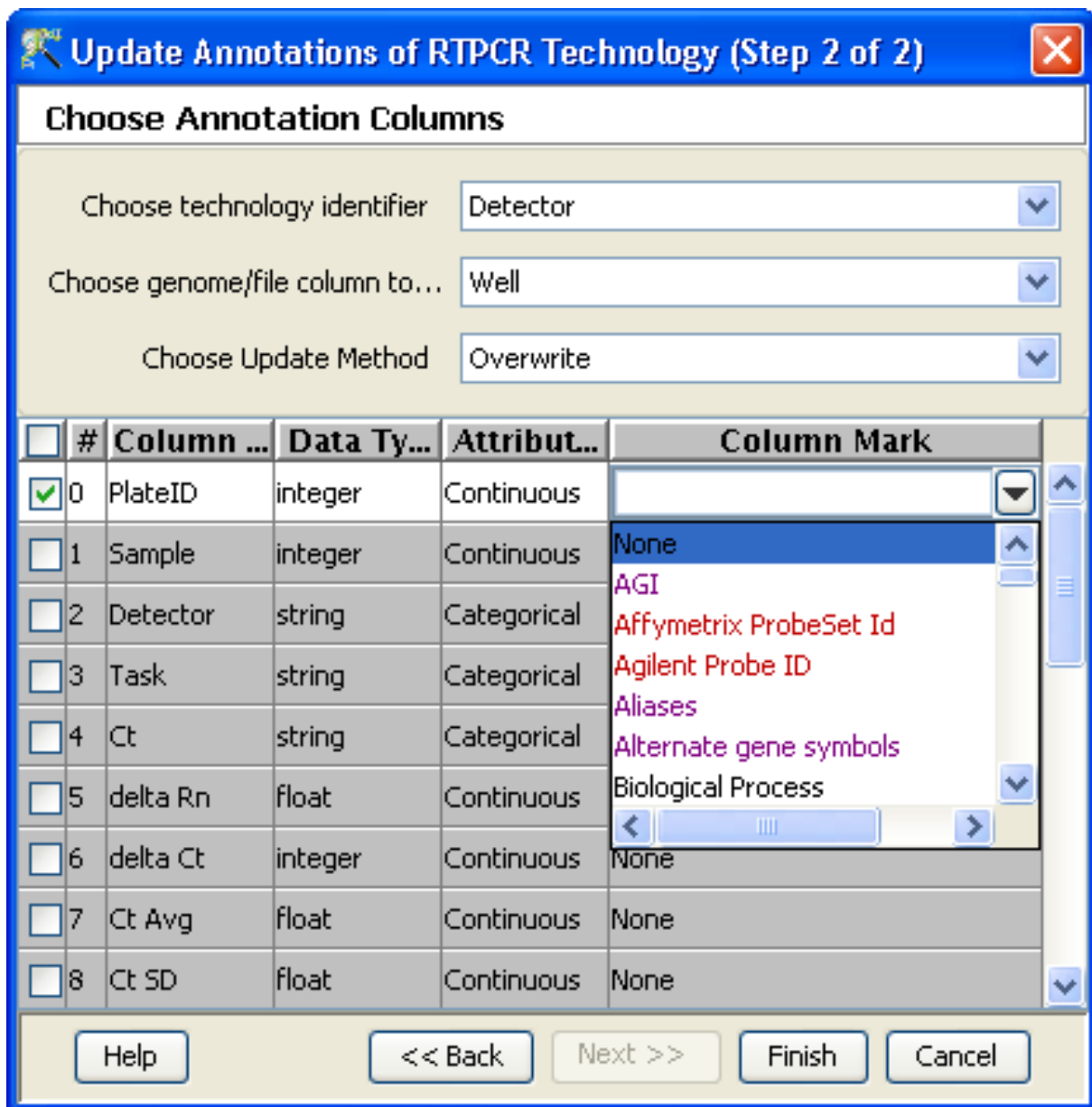


Figure 14.5: Choose Annotation Columns

Chapter 15

Analyzing Generic Single Color Expression Data

GeneSpring GX supports Generic Single Color technology. Any custom array with single color technology can be analyzed here. However, a technology first needs to be created, based upon the file format being imported.

15.1 Creating Technology

Technology creation is a step common to both Generic Single Color and Two color experiments. Technology creation enables the user to specify the columns (Signals, Flags, Annotations etc.) in the data file and their configurations which are to be imported. Different technologies need to be created for different file formats. Custom technology can be created by navigating to *Annotations* in the menu bar and selecting *Create Technology* → *Custom from file*. The process uses one data file as a sample file to mark the columns. Therefore, it is important that all the data files being used to create an experiment should have identical formats.

The *Create Custom Technology* wizard has multiple steps. While steps 1, 2, 3 and 9 are common to both the Single color and Two Color, the remaining steps are specific to either of the two technologies.

- (Step 1 of 9)
User input details, i.e., Technology type, Technology name, Organism, Sample data file location, Number of samples in a single data file and the Annotation file location are specified here. Files with a single sample or with multiple samples can be used to create the technology. Click *Next*. See Figure 15.1
- (Step 2 of 9) This allows the user to specify the data file format. For this operation, four options are

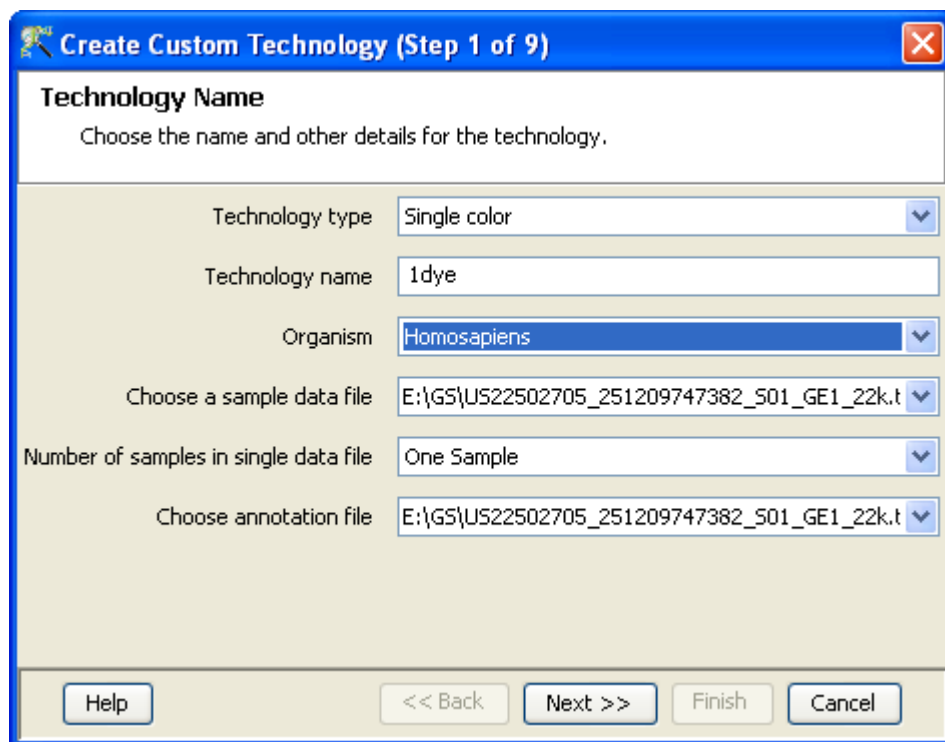


Figure 15.1: Technology Name

provided, namely, the *Separator*, the *Text qualifier*, the *Missing Value Indicator* and the *Comment Indicator*. The *Separator* option specifies if the fields in the file to be imported are separated by tab, comma, space etc. New separators can be defined by scrolling down to 'Enter New' and providing the appropriate symbol in the textbox. *Text qualifier* is used for indicating characters used to delineate full text strings. This is typically a single or double quote character. The *Missing Value Indicator* is for declaring a string that is used whenever a value is missing. This applies only to cases where the value is represented explicitly by a symbol such as N/A or NA. The *Comment Indicator* specifies a symbol or string that indicates a comment section in the input file. Comment Indicators are markers at the beginning of the line which indicate that the line should be skipped (typical examples is the # symbol). See Figure 15.2

- (Step 3 of 9) The data files typically contain headers which are descriptive of the chip type and are not needed for the analysis. Only those rows containing the data values are required. The purpose of this step is to identify which rows need to be imported. The rows to be imported must be contiguous in the file. The rules defined for importing rows from this file will then apply to all other files to be imported using this technology. Three options are provided for selecting rows: The default option is to select all rows in the file. Alternatively, one can choose to take a block of rows between specific row numbers (use the preview window to identify row numbers) by entering the row numbers in the appropriate textboxes. Remember to press the Enter key before proceeding. In addition, for situations where the data of interest lies between specific text markers, those text markers can be indicated. Note also that instead of choosing one of the options from the radio buttons, one can choose to select specific contiguous rows from the preview window itself by using Left-Click and Shift-Left-Click on the row header. The panel at the bottom should be used to indicate whether or

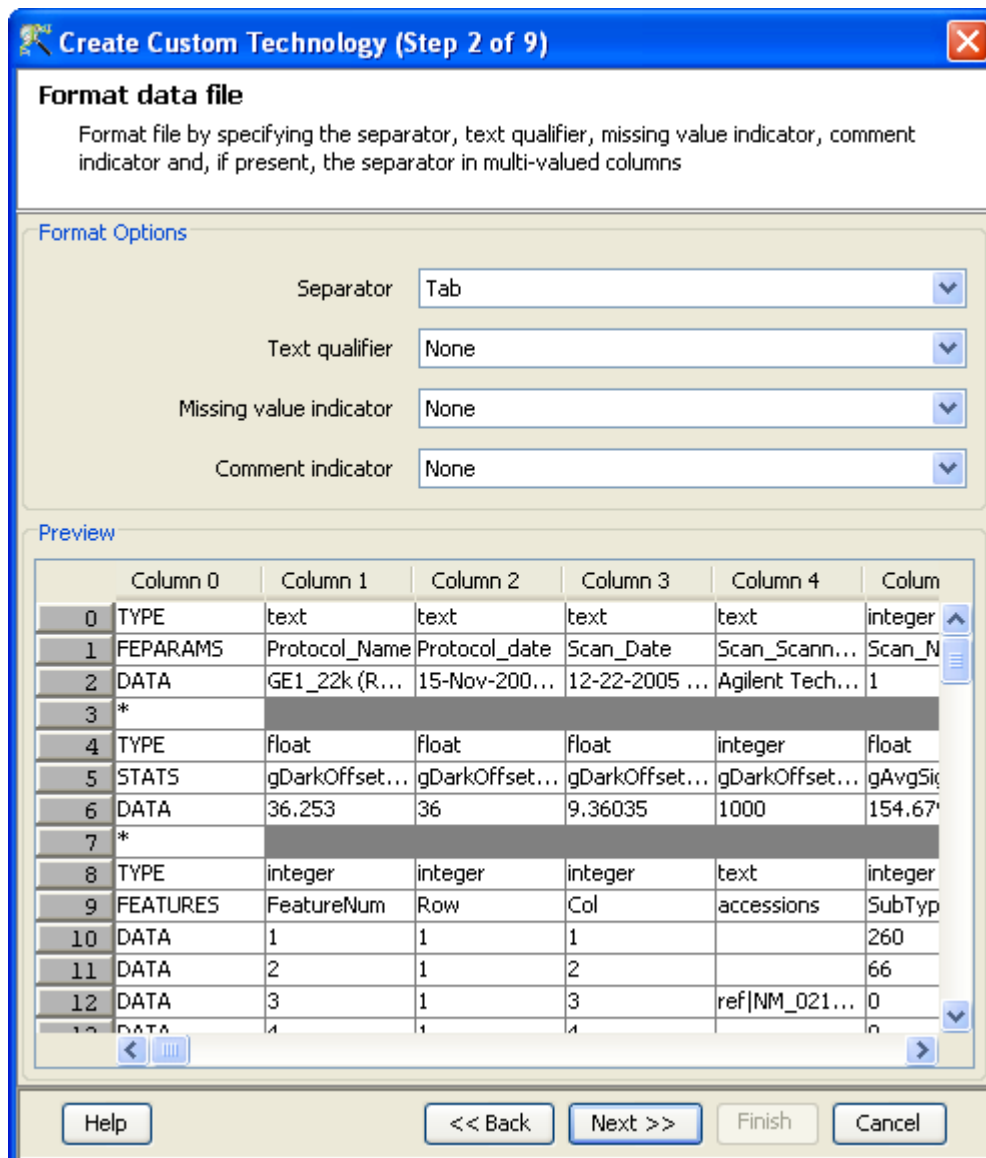


Figure 15.2: Format data file

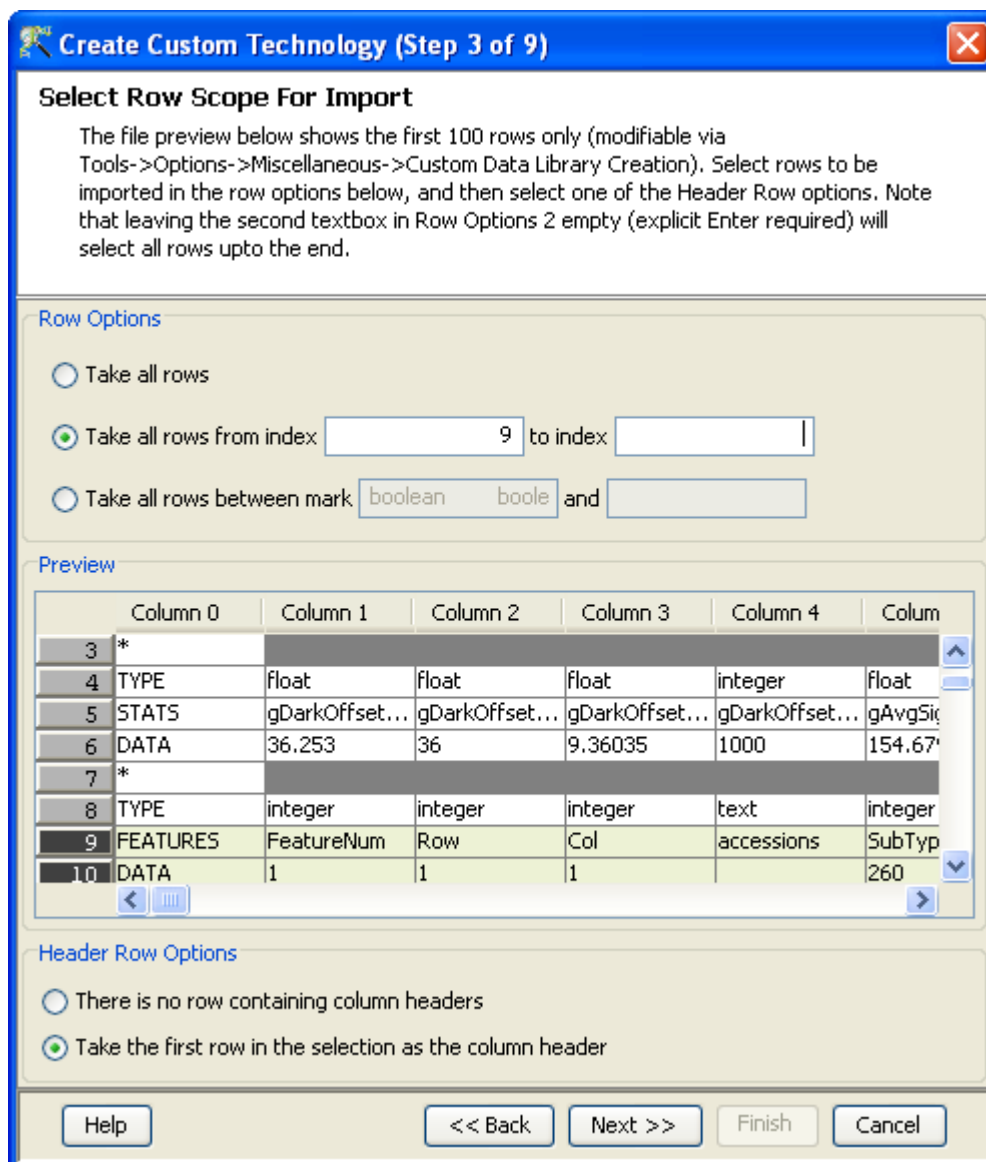


Figure 15.3: Select Row Scope for Import

not there is a header row; in the latter case, dummy column names will be assigned. See Figure 15.3

- (Step 4 of 9) This step is specific for file formats which contain a single sample per file. Gene identifier, background(BG) corrected signal and the flag columns are indicated here. Flag column can be configured using the *Configure* button to designate Present(P), Absent(A) or Marginal(M) values. See Figure 15.4
- (Step 5 of 9)

This step is specific for file formats which contain multiple samples per file. Such file formats typically contain a single column having the identifier and multiple columns representing the samples (one data column per sample). In this step, the Identifier column has to be indicated. The signal and flag columns for each sample also should be identified here and moved from *All columns* to *Signal columns*

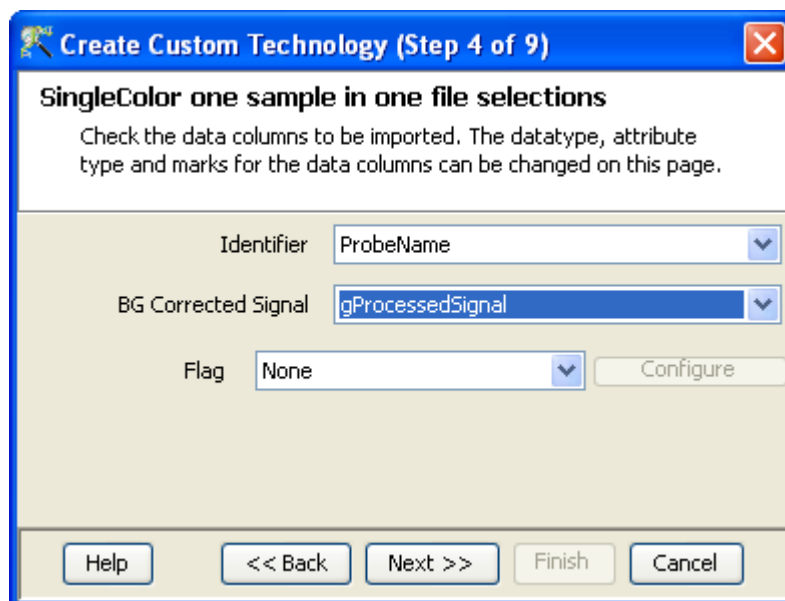


Figure 15.4: Single Color one sample in one file selections

and *Flag columns* box respectively. This can be done either by putting in the *Keyword* for the Signal and the Flag columns and clicking *Refresh* or by choosing *Custom* and selecting the columns as either Signal or Flag in the *Identify signal and flag columns by* option. After marking 2 columns, the user can utilize the option *Guess the Rest* for marking the other columns. The tool tries to match the names of the selected columns with the rest and marks those columns which have similar names to the selected ones. The *Choose representative flag* allows the user to choose one of the flag columns for configuring the flag settings. See Figures 15.5 and 15.6.

- (Steps 6 of 9)
This step of the wizard is used in case of technology creation for 2-dye or 2-color samples.
- (Steps 7 of 9)
This step is similar to the step 2 of 9 and is used to format the annotation file. If a separate annotation file does not exist, then the same data file can be used as an annotation file, provided it has the annotation columns.
- (Step 8 of 9) Identical to step 3 of 9, this allows the user to select row scope for import in the annotation file.
- (Step 9 of 9) The Step 9 of technology creation is an extremely important step which allows the user to mark the columns appropriately. Proper marking of the various columns in the annotation file will enable the various functionalities like GO, GSEA, Genome Browser, Pathway Analysis to proceed smoothly. The markings to be given for all these functions are elaborated below:
- GSEA: The annotation file should contain a column containing the Gene Symbol. This column should be marked as *Gene Symbol* from the drop-down menu.

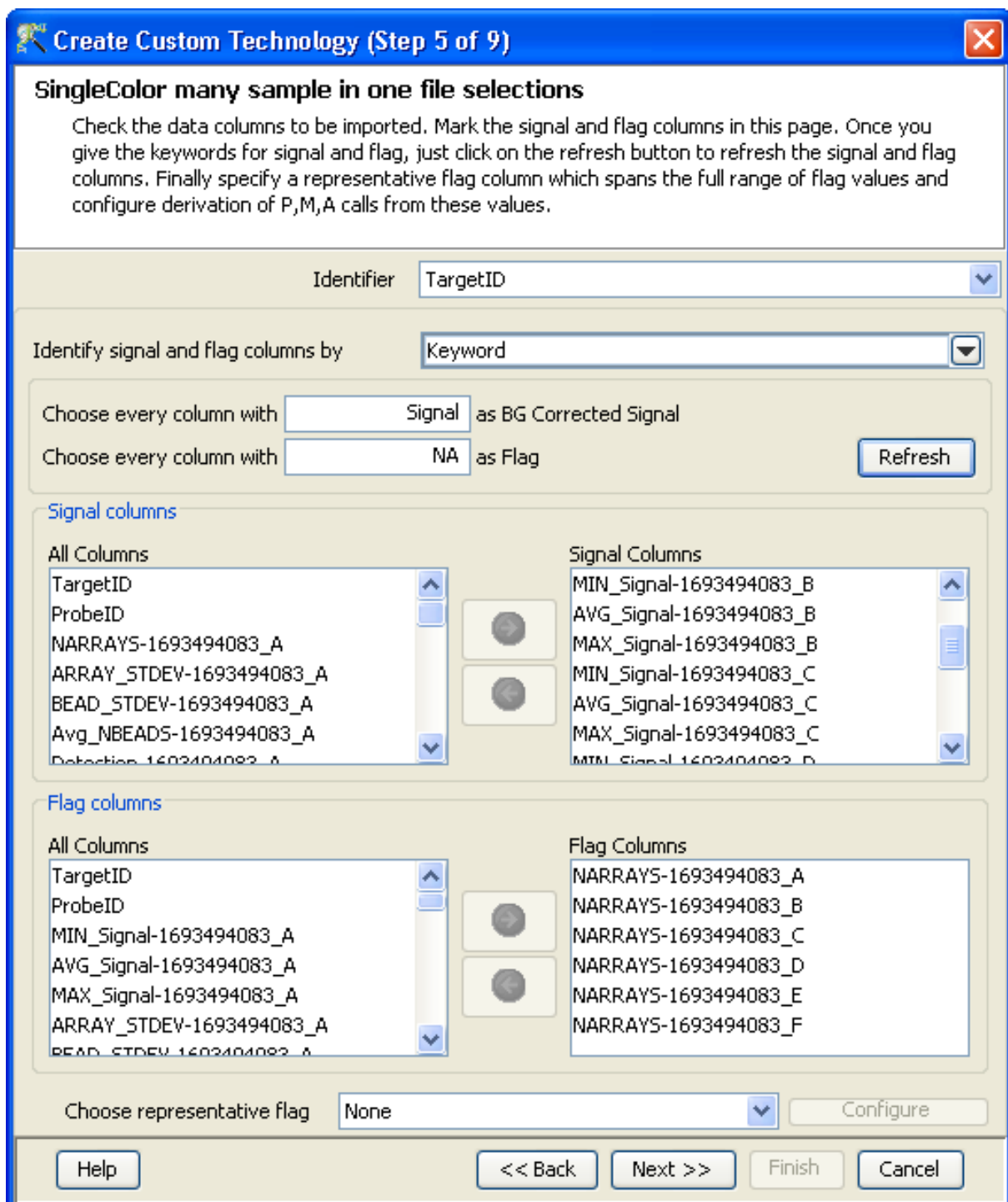


Figure 15.5: Single Color-Multiple Samples Per File-Keyword Selection

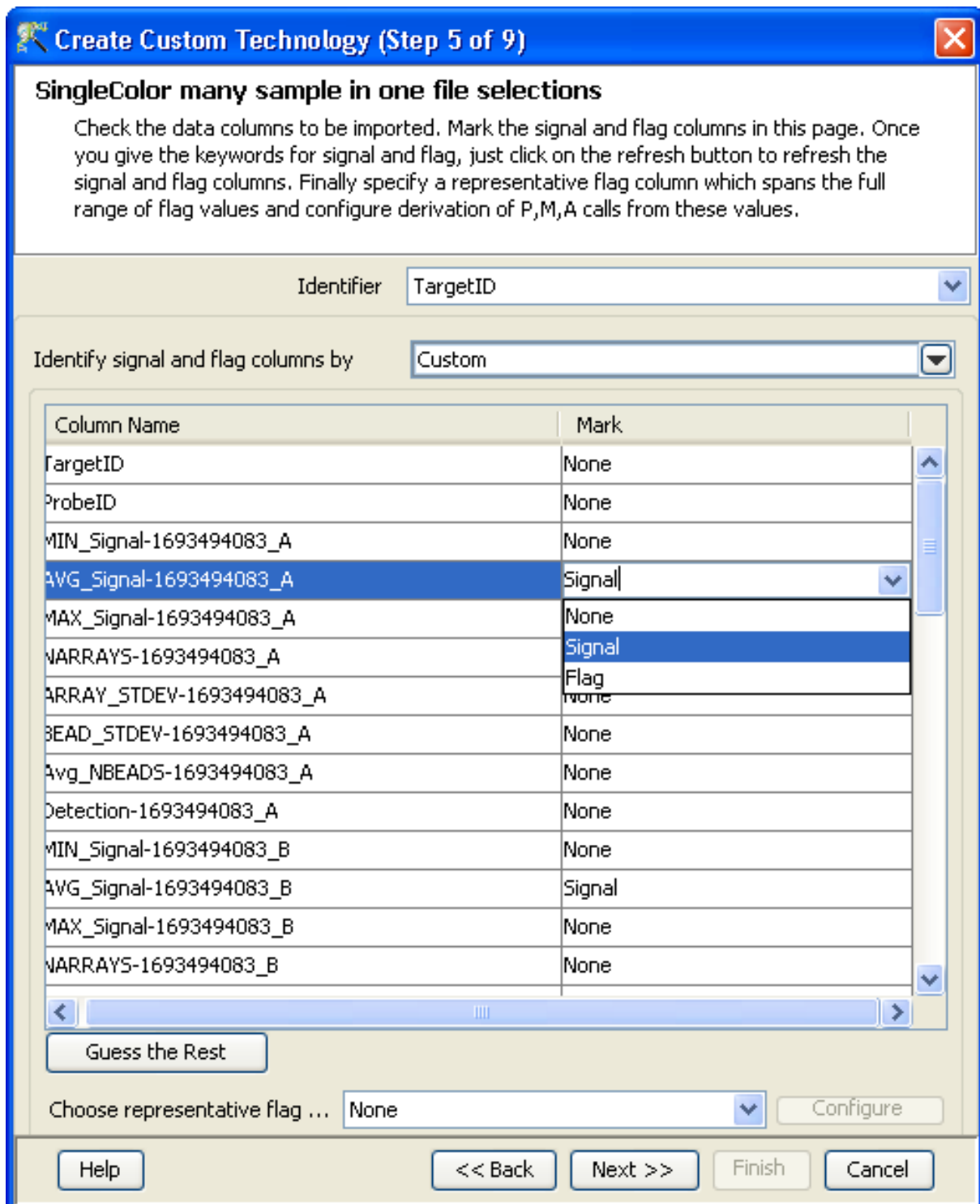


Figure 15.6: Single Color-Multiple Samples Per File-Custom Selection

- GSA: The annotation file should contain a column containing the Gene Symbol. This column should be marked as *Gene Symbol* from the drop-down menu.
- GO: For carrying out GO analysis, the annotation file can either contain a single column with all the GO IDs in it, separated by a separator or it can contain separate columns for the different GO processes. The single column with multiple GO IDs should be marked as *Gene Ontology accession* from the drop-down menu. Instead if columns containing individual GO processes(Biological Process, Cellular Component and Molecular Function) are present, they should be marked accordingly in the dropdown menu.
- Genome Browser: In order to view the data in Genome Browser, the annotation file should contain a Chromosome Start Index, Chromosome End Index, Strand and Chromosome Number columns. Provide the column mark for Chromosome Start index, Chromosome End index, Strand, Chromosome number respectively, from the drop-down menu.

Note: The Chromosome Start index < Chromosome End index. For viewing Profile track only, in the Genome Browser, chromosome start index and chromosome number are needed. The labelling of the chromosome numbers should follow this convention-chr1, chr2i.e. the word starts with chr followed by the chromosome number (without any space). For viewing data track, all four Chromosome Start index, Chromosome End index, Strand, Chromosome number are needed.

- If a custom technology is being created using an **Illumina** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

For viewing using the Genome Browser, the annotation files has three columns which have values for all four (Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand). Therefore before creating the custom experiment the user needs to parse these columns and create three new columns as follows :

Probe_Chr_Orientation- This column can be taken as it is. It should be marked as Strand.

Chromosome - A new column must be created wherein a 'chr' should be appended to each entry in the Chromosome column and this new column should be marked as Chromosome Number.

Probe_Coordinates- This column has each entry in the format a-b where a < b. Two new columns need to be created. one which has only the a values, (it should be marked as Chromosome Start Index) one which has only the b values (it should be marked as Chromosome End Index).

- If a custom technology is being created using an **Agilent** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

The annotation files have a single column 'Map' which has values for all four Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand. Therefore before creating the custom experiment the user needs to parse the file and separate the four columns as Chromosome Start Index, Chromosome End Index Chromosome Number and Strand.

Each entry in the Map column is typically in the format chrQ:a..b

if $a < b$, the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is a; the corresponding Chromosome End Index is b; the corresponding Strand is + .

if $a > b$ the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is b; the corresponding Chromosome End Index is a; the corresponding Strand is - .

For example, a Map value of chr14:34101457..34101398 corresponds to a Chromosome Start Index of 34101398, a Chromosome End Index of 34101457, a Chromosome Number of chr14 and a Strand of - (because in chrX:a..b $a > b$)

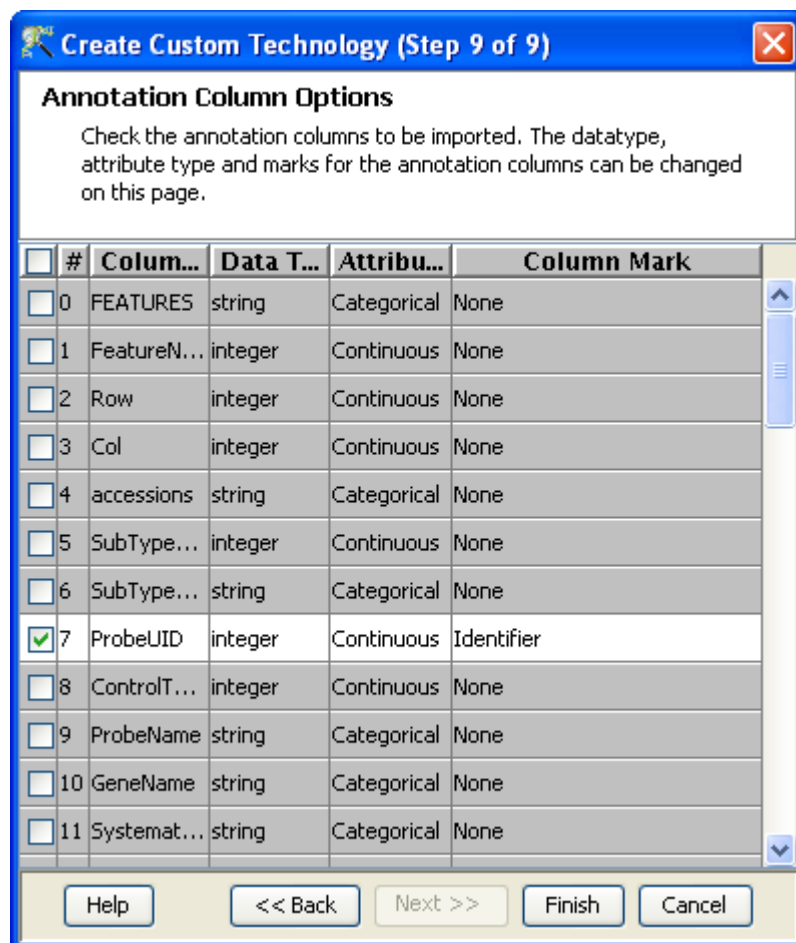


Figure 15.7: Annotation Column Options

For example, a Map value of chr6:46222041..46222100 corresponds to a Chromosome Start Index of 46222041, a Chromosome End Index of 46222100, a Chromosome Number of chr6 and a Strand of +(because in chrX:a..b a<b)

- Import BioPAX pathways: Pathways being imported should be in .owl format. During custom technology creation, provide the column mark for Entrez Gene ID/SwissProt from the drop-down menu. Only after this mark is provided can the proteins involved in a particular pathway be highlighted.
- Find Similar Pathways: The annotation file should contain an Entrez Gene ID/SwissProt column, which have to be marked appropriately as Entrez Gene ID/SwissProt.
- Translation: This operation can be performed between organisms listed in the **Homologene table** in section [Translation](#). Entrez Gene ID column has to be marked for performing translation.

See Figure [15.7](#)

The types of Data and Attribute marks available for the annotation columns are

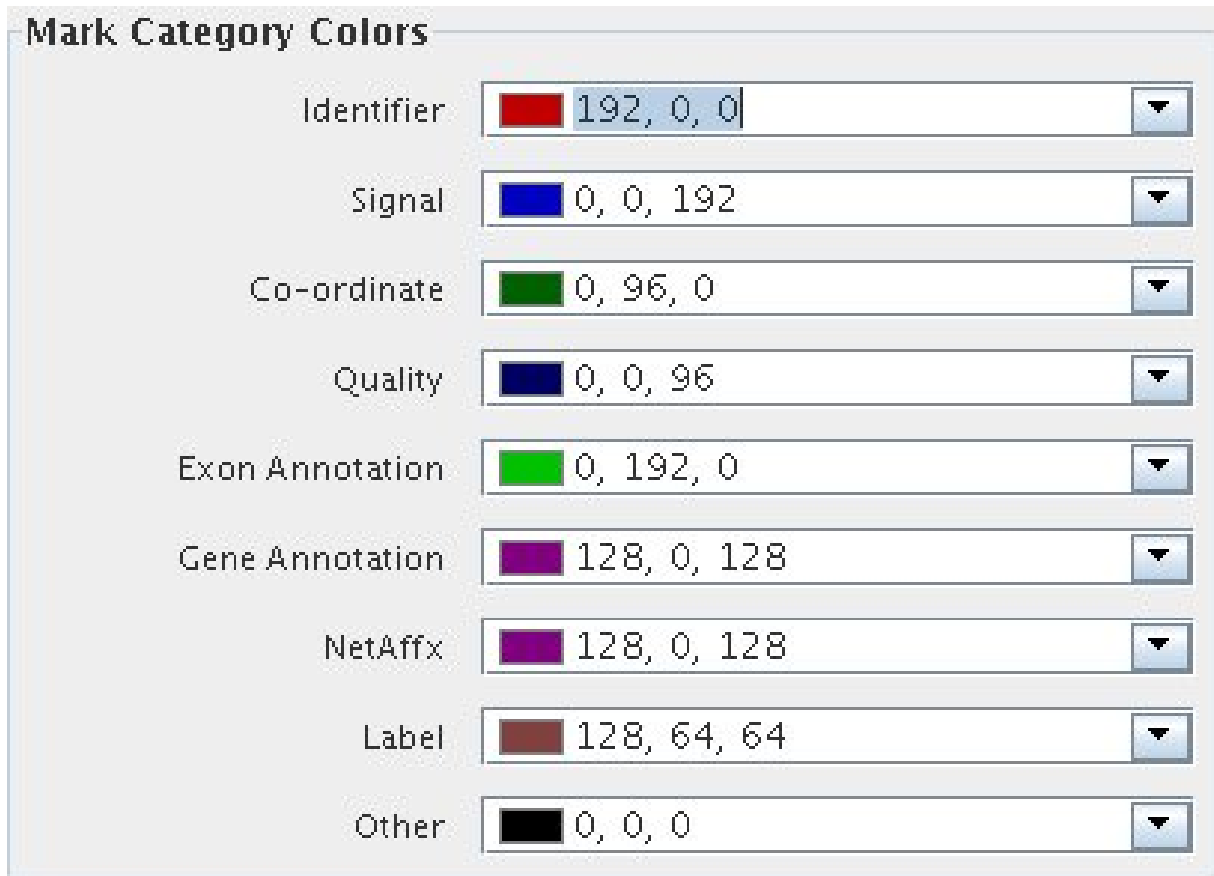


Figure 15.8: Annotation Mark Colors

- **Categorical:** A column marked as a "categorical" column means that the values in this column fall into certain finite distinct categories.
- **Continuous:** A column marked as a "continuous" column means that the values in this column can vary, potentially, over any large range.
- **String:** A continuous sequence of symbols or digits, not including a space.
- **Float:** A real number, i.e a number which can be given by a decimal representation.

The annotation marks are colored on the basis of their functionality in the tool. The meaning of the various colors are provided in the figure 16.5. This figure is provided solely for visualization purposes and is not available from the tool.

Click **Finish** to exit the wizard.



Figure 15.9: Welcome Screen

15.1.1 Project and Experiment Creation

After technology creation, data files satisfying the file format can be used to create an experiment. The following steps will guide you through the process of experiment creation.

Upon launching **GeneSpring GX**, the startup is displayed with 3 options.

1. **Create new project**
2. **Open existing project**
3. **Open recent project.**

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting *Create New Project*, a window appears in which details (name of the project and notes) can be recorded. Press *OK* to proceed.

An Experiment Selection Dialog window then appears with two options.

1. **Create new experiment**
2. **Open existing experiment**

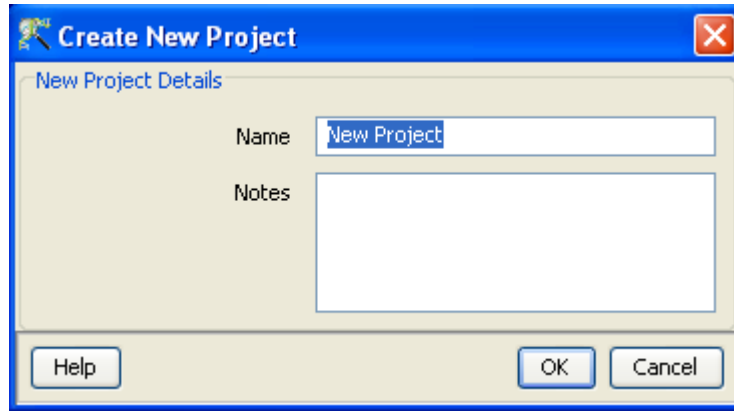


Figure 15.10: Create New project

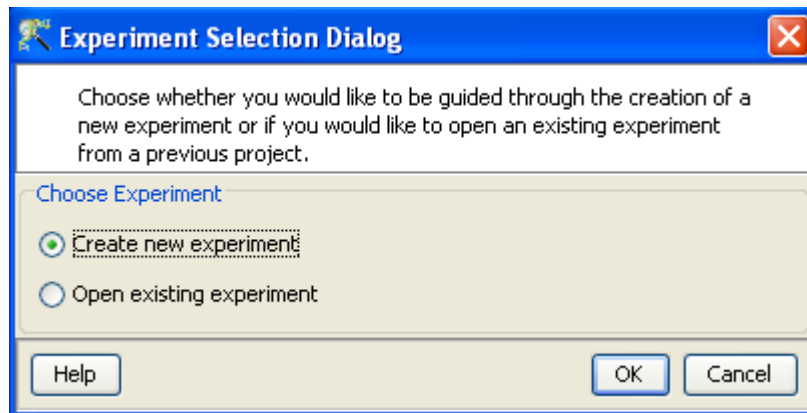


Figure 15.11: Experiment Selection

Selecting *Create new experiment* allows the user to create a new experiment (steps described below). *Open existing experiment* allows the user to use existing experiments from any previous projects in the current project. Choosing *Create new experiment* opens up a New Experiment dialog in which *Experiment name* can be assigned. The *Experiment type* should then be specified (Generic Single Color), using the drop down button. The *Workflow Type* can be used to choose whether the workflow will be *Guided* or *Advanced*. Unlike the other technologies where *Guided* and *Advanced* analysis workflows are available, in case of Generic Single Color, only the *Advanced Workflow* is supported. Click *OK* will open a new experiment wizard. See Figure 15.12

15.2 Data Processing for Generic Single Color Experiment

1. **File formats:** The files should be tabular in nature. For example, .csv, .tsv, .txt etc. can be used.
2. **Raw:** The term "raw" signal values refer to the linear data after thresholding and summarization. Summarization is performed by computing the geometric mean.

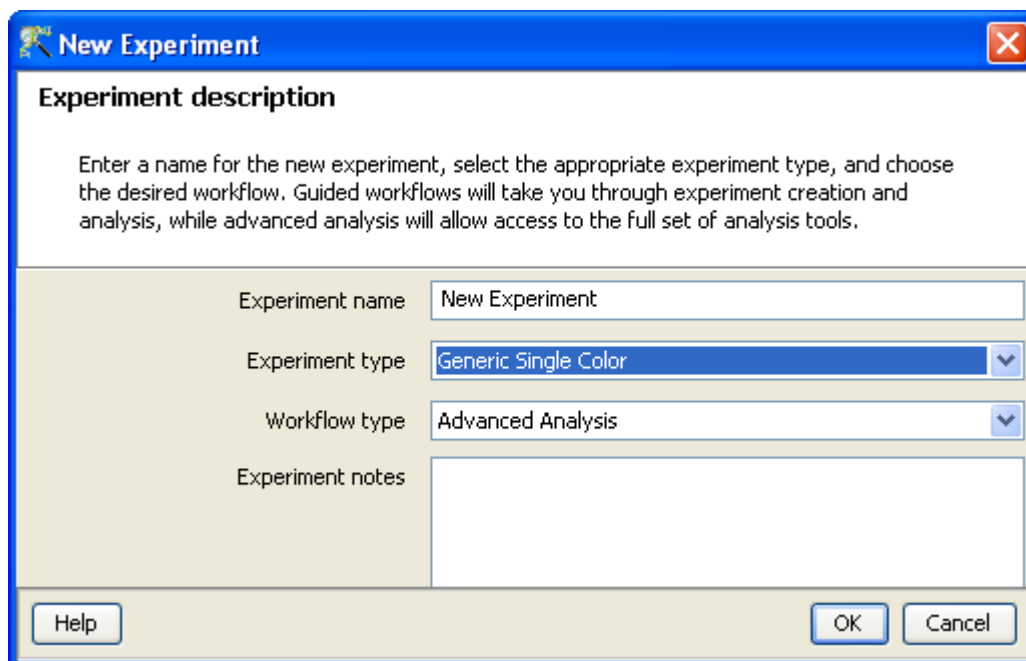


Figure 15.12: Experiment Description

3. **Normalized:** "Normalized" value is the value generated after log transformation and normalization (Percentile Shift, Scale, Normalize to control genes or Quantile) and Baseline Transformation.
4. **Treatment of on-chip replicates:** The signal value of a probeset is the geometric mean of all its probes.
5. **Flag values:** The values for the probes are configured by the user during the creation of technology as either Present, Marginal or Absent. Based on the values of the probes, the probeset is assigned a flag value. The order of importance for flag values for probes in a probeset is *Present*>*Marginal*>*Absent*.
6. **Treatment of Control probes:** The control probes are included while performing normalization.
7. **Empty Cells:** Empty cells might be present in the intensity values column for certain genes in the data file. These genes are brought in **GeneSpring GX** . But an entity list containing these genes cannot be used for running Clustering and Class Prediction analyses. The user can choose to remove the missing values from an entity list using the option *Remove Entities with missing signal values* from the *Results Interpretations* section of the workflow.
8. **Sequence of events:** The sequence of events involved in the processing of the data files is: thresholding, summarization, log transformation and Normalization followed by Baseline Transformation.
9. **Merging of files:** Multiple files in Generic experiment creation are combined based on the Identifier column using the following rules. The very first file among the various files chosen server as a master reference (you can determine which file serves as the first file using the *Reorder* button on Page 1 of the New Experiment Creation page). The number of rows in this master must exceed the number of rows in all subsequent files, for extra rows in these subsequent files are dropped. Next, all identifiers in the Identifier column of this first file are considered and missing values in these, if any, are discarded.

This results in a set of valid identifier values; all rows in all other files whose identifier values are outside of this set are discarded. Next, on-chip replicates are determined by counting the number of occurrences of each valid identifier in the first file. Consider for example an identifier Id1 which appears 3 times in file 1. Then rows corresponding to the first 3 occurrences of Id1 are taken in each of the other files; if there are fewer than 3 rows, then as many rows that are present are taken; and if there are more than 3 rows, then the first 3 are taken. The summarized value for Id1 in each file is determined by taking a geometric mean over these chosen rows.

15.3 Advanced Analysis

The **Advanced Workflow** offers a variety of choices to the user for the analysis. Raw signal thresholding can be altered. Based upon the technology, Quantile or Median Shift normalization can be performed. Additionally there are options for baseline transformation of the data and for creating different interpretations. To create and analyze an experiment using the **Advanced Workflow**, choose the **Workflow Type** as Advanced. Clicking **OK** will open a New Experiment Wizard, which then proceeds as follows:

1. New Experiment (Step 1 of 4): The technology (created as mentioned above) can be selected and the new data files or previously used data files in **GeneSpring GX** can be imported in to create the experiment. A window appears containing the following options:
 - (a) **Choose Files(s)**
 - (b) **Choose Samples**
 - (c) **Choose Raw Files**
 - (d) **Reorder**
 - (e) **Remove**

An experiment can be created using either the data files or else using samples. Upon loading data files, **GeneSpring GX** associates the files with the technology (see below) and creates samples. These samples are stored in the system and can be used to create another experiment via the **Choose Samples** option through a sample search wizard. If the user has imported any custom experiments from **GeneSpring GX 7** and wants to recreate the experiment in **GeneSpring GX**, then the user can create a new technology in the tool with an original raw file and later utilize the **Choose Raw Files** option to choose the raw files associated with the migrated custom experiment. For selecting data files and creating an experiment, click on the **Choose File(s)** button, navigate to the appropriate folder and select the files of interest. The files can be either tab separated (.txt or .tsv) or could be comma separated (.csv). Select **OK** to proceed.

The sample search wizard that comes up via the option *Choose Samples* has the following search conditions:

- (a) **Search field** (which searches using any of the 6 following parameters- (Creation date, Modified date, Name, Owner, Technology, Type).
- (b) **Condition** (which requires any of the 4 parameters- (equals, starts with, ends with and includes Search value).

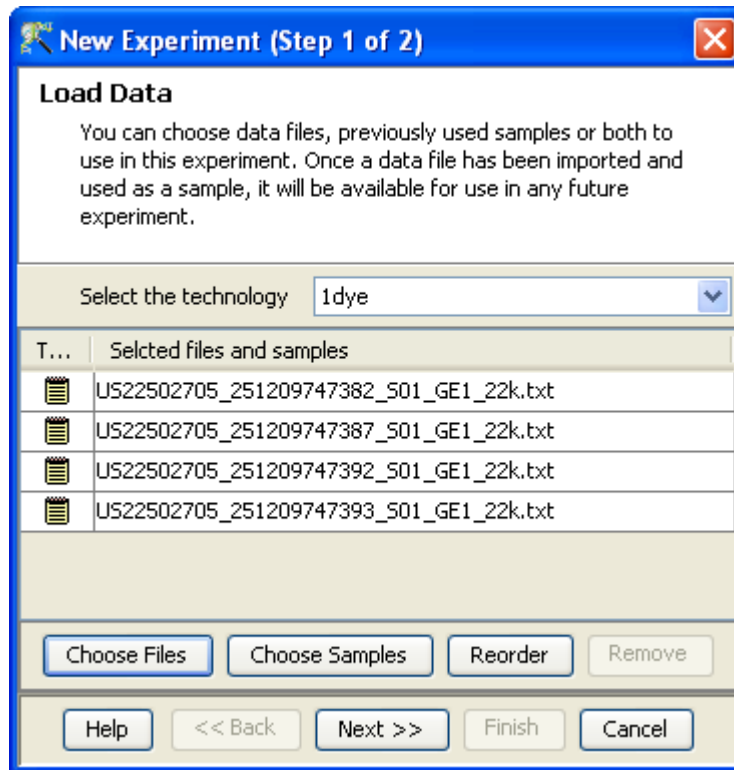


Figure 15.13: Load Data

(c) **Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment using *Add* button, similarly can be removed using *Remove* button.

After selecting the files, clicking on the *Reorder* button opens a window in which the particular sample or file can be selected and can be moved either up or down by pressing on the buttons. Click on *OK* to enable the reordering or on *Cancel* to revert to the old order. See Figure 15.13

2. New Experiment (Step 2 of 4): This gives the options for preprocessing of input data. It allows the user to threshold raw signals to chosen values and to select normalization algorithms(Quantile, Percentile Shift, Scale and Normalize to control genes).

- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at

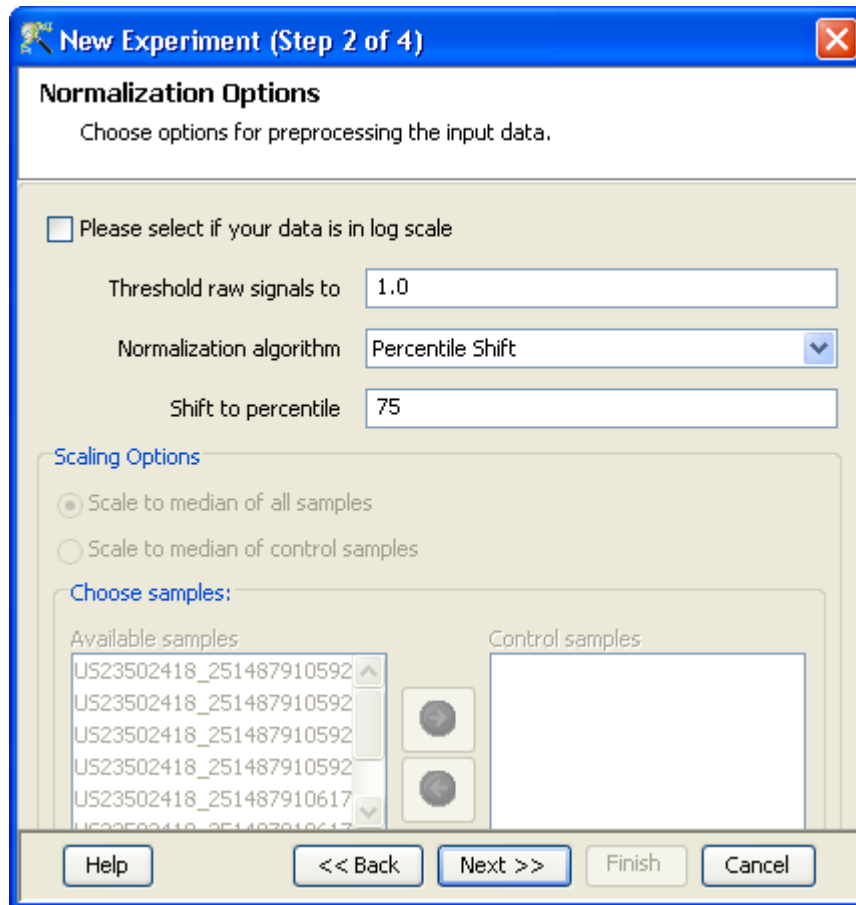


Figure 15.14: Preprocess Options

a default value of 50.

- **Normalize to External Value:** This option will bring up a table listing all samples and a default scaling factor of '1.0' against each of them. The user can use the '*Assign Value*' button at the bottom to assign a different scaling factor to each of the sample; multiple samples can be chosen simultaneously and assigned a value.

For details on the above normalization methods, refer to section [Normalization Algorithms](#).

In case, the data is already log transformed, the user can select the checkbox stating that their signal values are already in log scale. This will disable the thresholding option also.

See figure [15.14](#).

Experiment (Step 3 of 4): If the **Normalize to control genes** option is chosen, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the

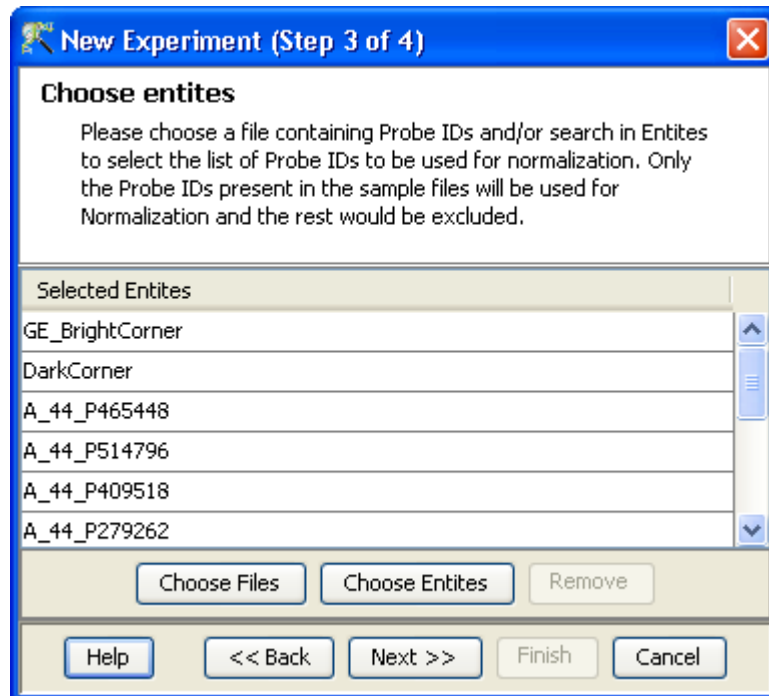


Figure 15.15: Choose Entities

technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selection is over. See figure 15.15.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe ids whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Experiment (Step 4 of 4): This step allows the user to perform baseline transformation. See figure 15.16. The baseline options include

- *Do not perform baseline*
- *Baseline to median of all samples:* For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- *Baseline to median of control samples:* For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on *Assign value*. This opens up the *Choose Control Samples* window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on *Ok*. This will show the control samples for each of the samples.

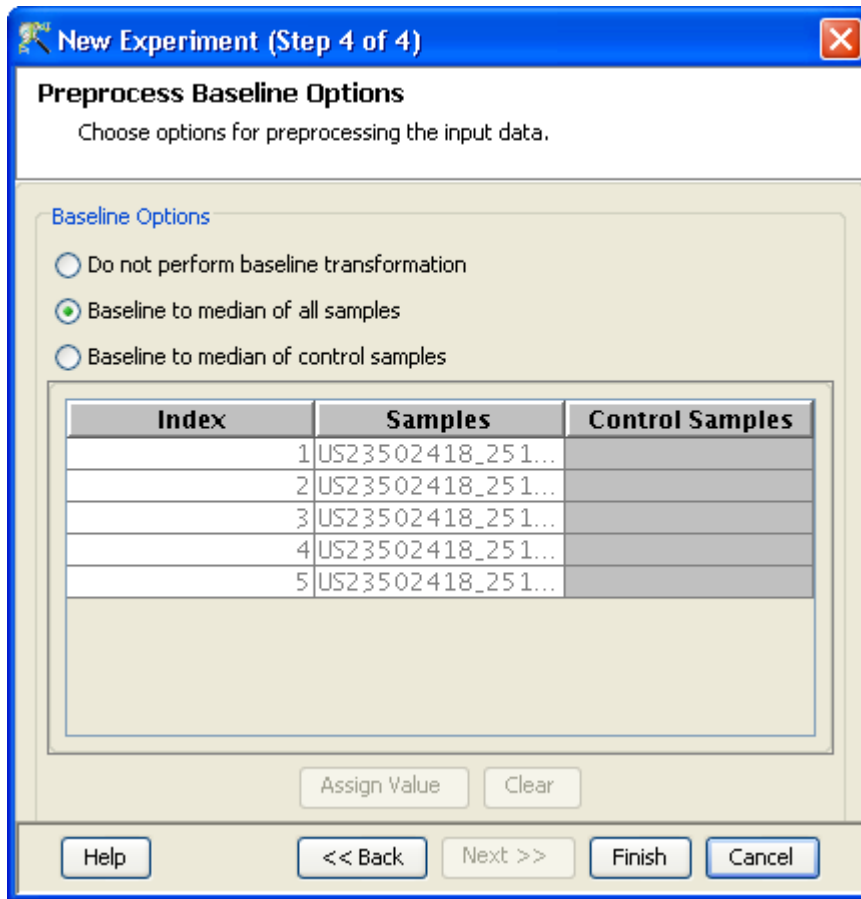


Figure 15.16: Preprocess Baseline Options

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

Clicking **Finish** creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to **View** in Toolbar.

15.3.1 Experiment Setup

- *Quick Start Guide*: Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the Advanced workflow, the method of analysis, the details of the algorithms used and the interpretation of results
- *Experiment Grouping*: Experiment parameters defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)

- **Create Interpretation** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.

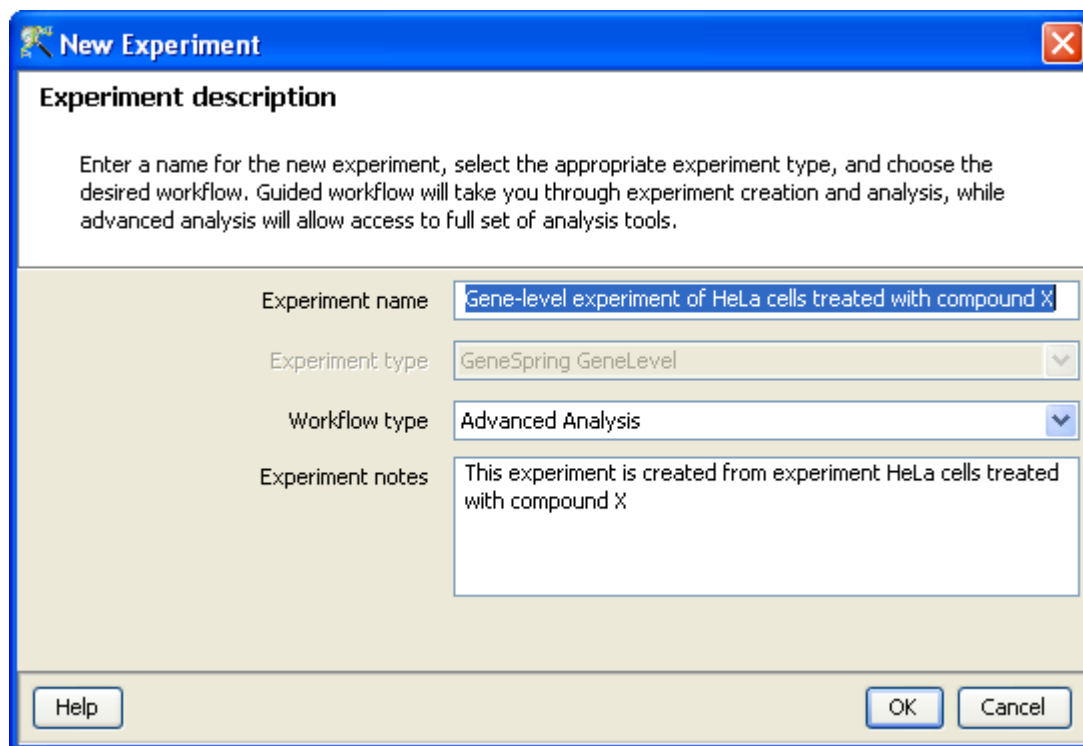


Figure 15.17: Gene Level Experiment Creation

- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the **Choose Entities** option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click **Finish**. This will result in the entities getting selected as control entities and will appear in the wizard.

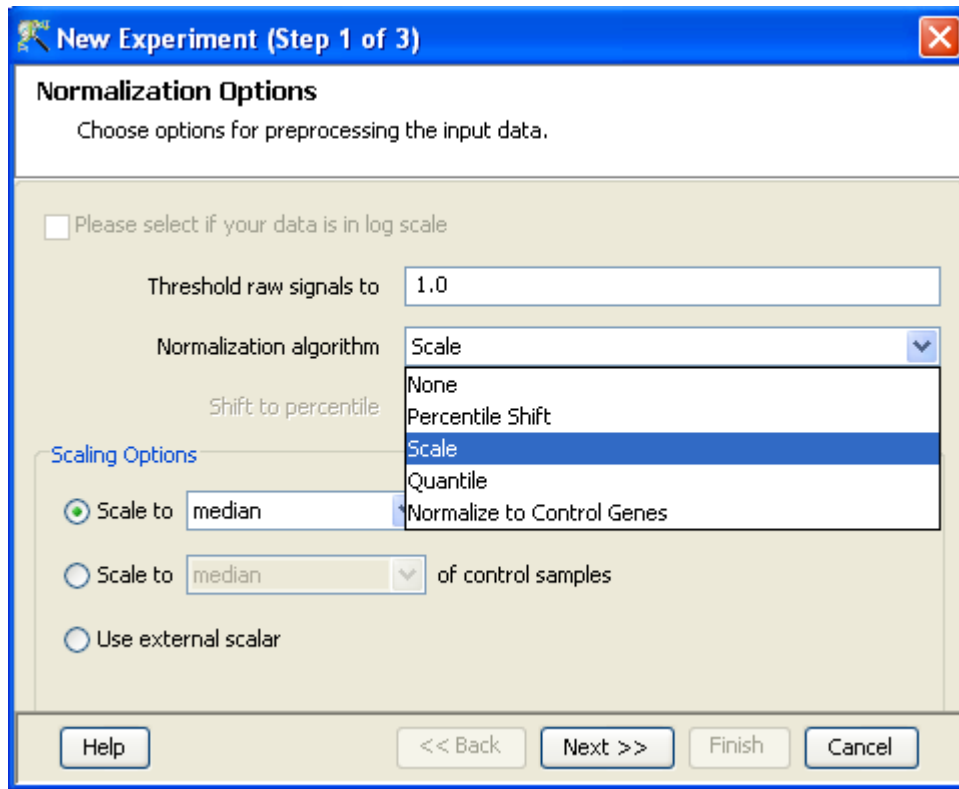


Figure 15.18: Gene Level Experiment Creation - Normalization Options

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

15.3.2 Quality Control

- **Quality Control on Samples:** The view shows four tiled windows
 1. Correlation coefficients table and Correlation coefficients plot tabs
 2. Experiment grouping
 3. PCA scores

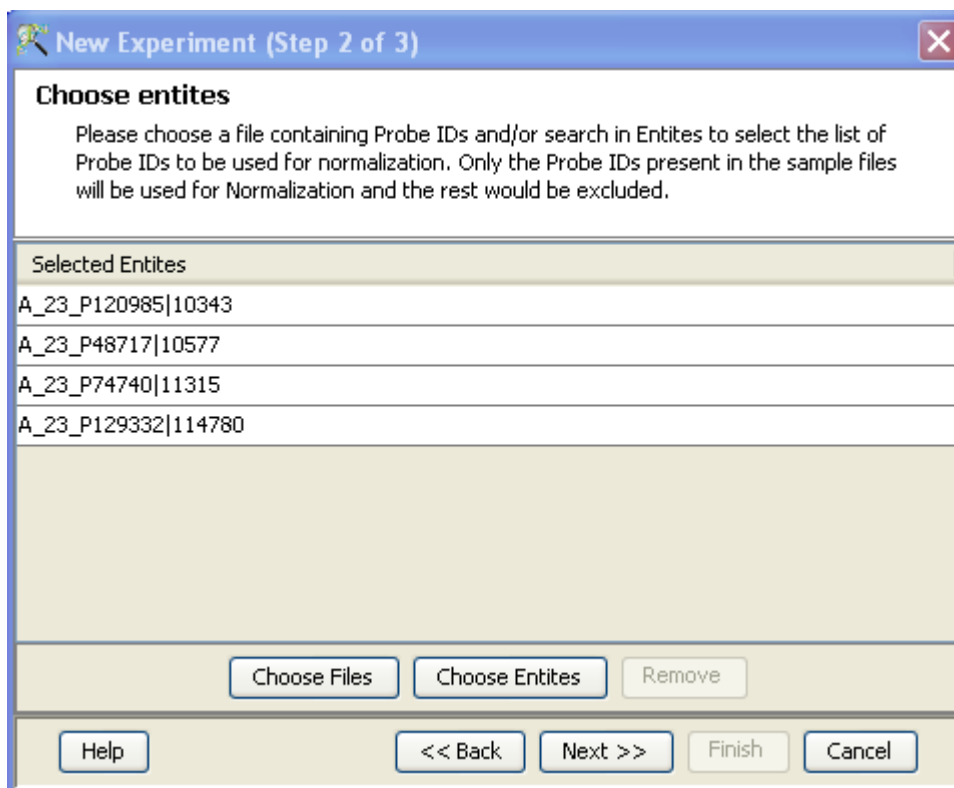


Figure 15.19: Gene Level Experiment Creation - Choose Entities

4. Legend

See Figure 15.21

The Correlation Plots shows the correlation analysis across arrays. It finds the correlation coefficient for each pair of arrays and then displays these in two forms, one in textual form as a correlation table view which also shows the experiment grouping information, and other in visual form as a heatmap. The correlation coefficient is calculated using Pearson Correlation Coefficient.

Pearson Correlation: Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$

The heatmap is colorable by Experiment Factor information via Right-Click→Properties. The intensity levels in the heatmap can also be customized here.

NOTE: The Correlation coefficient is computed on raw, unnormalized data and in linear scale. Also, the plot is limited to 100 samples, as it is a computationally intense operation.

Experiment Grouping shows the parameters and parameter values for each sample.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows

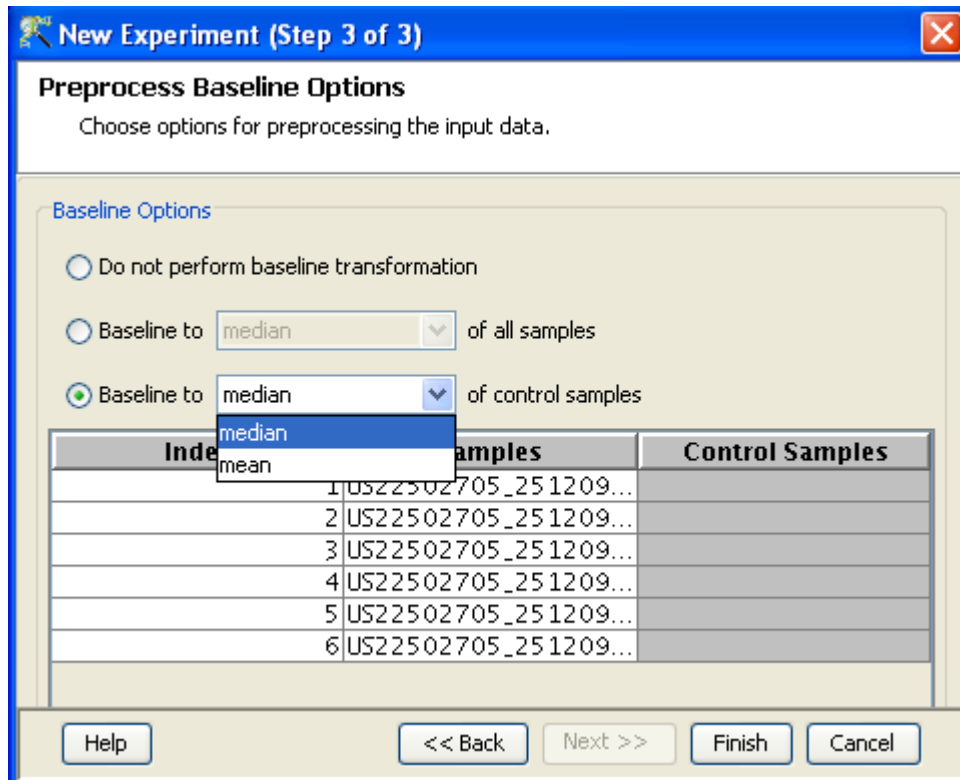


Figure 15.20: Gene Level Experiment Creation - Preprocess Baseline Options

viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

Click on **OK** to proceed.

- **Filter Probe Set by Expression:**

Entities are filtered based on their signal intensity values. For details refer to the section on [Filter Probesets by Expression](#)

- **Filter Probe Set by Flags:**

In this step, the entities are filtered based on their flag values P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new technology (Step 4 of 9) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window.

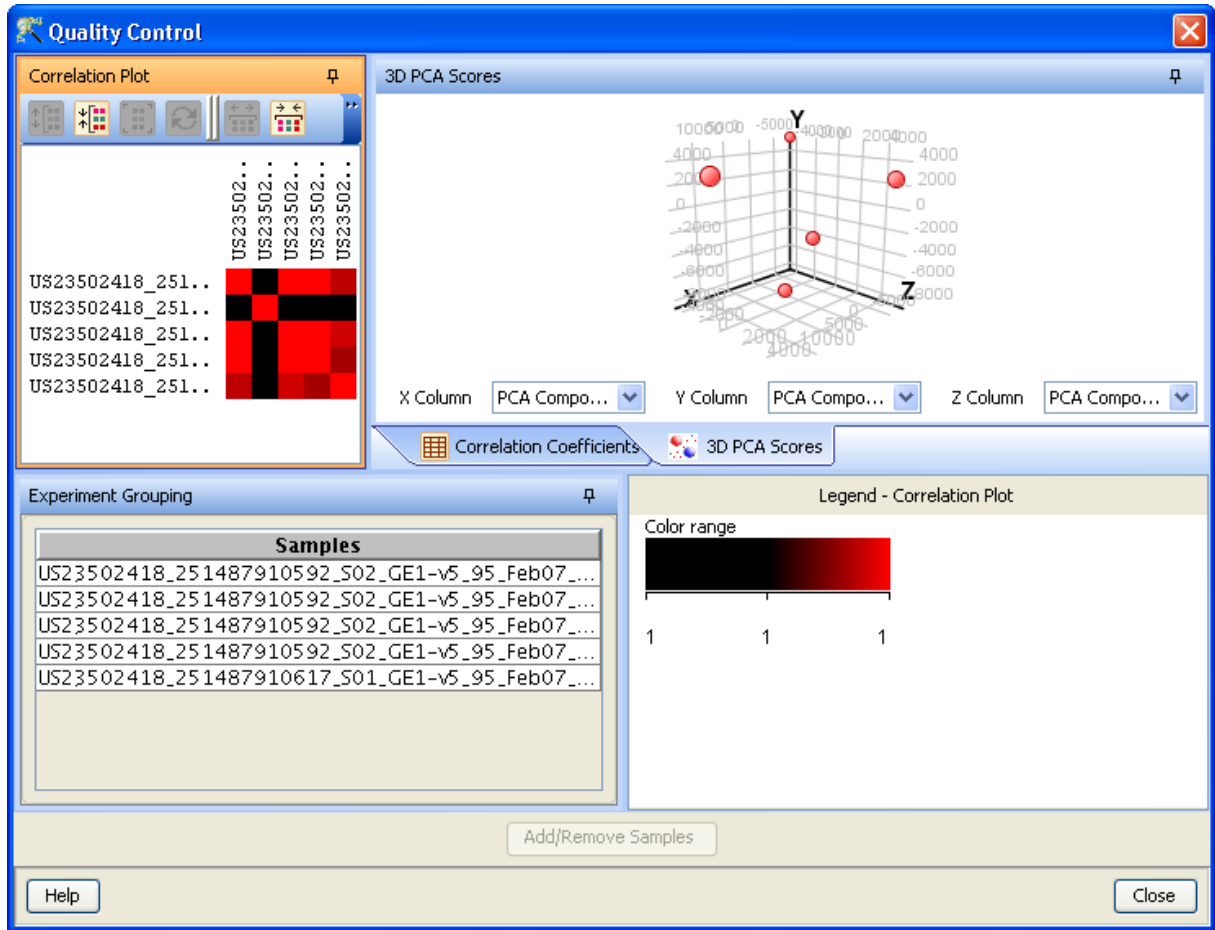


Figure 15.21: Quality Control

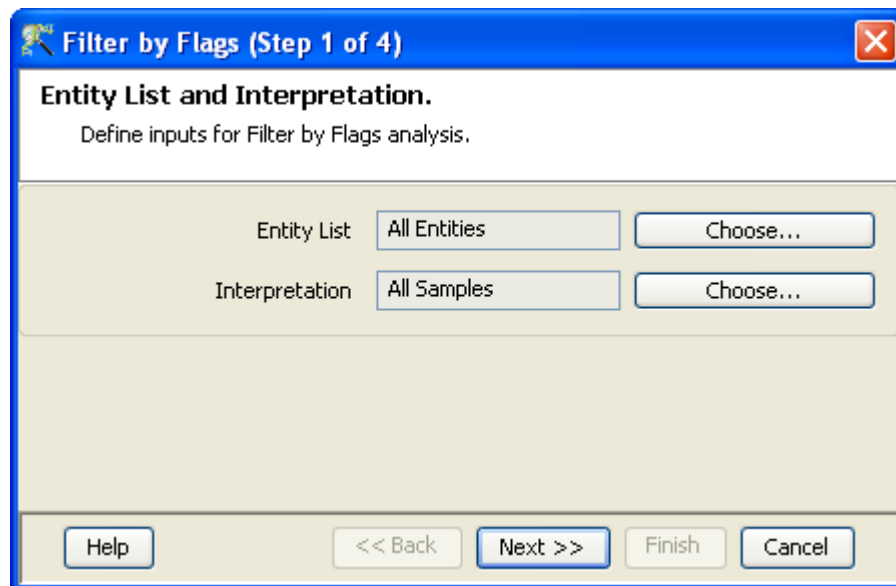


Figure 15.22: Entity list and Interpretation

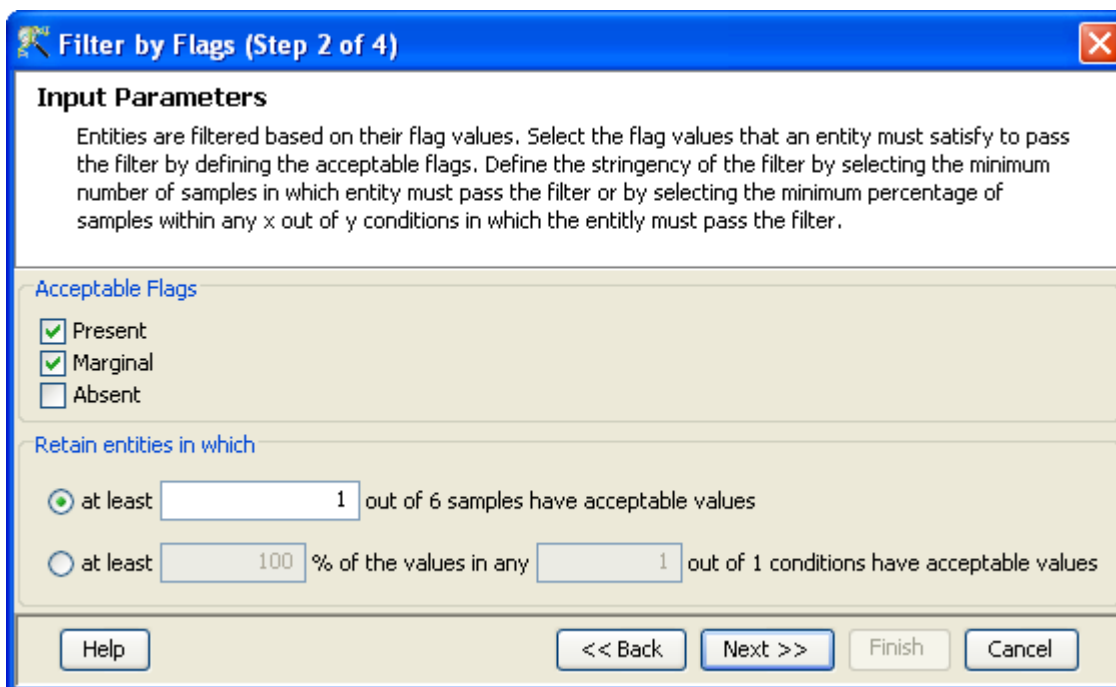


Figure 15.23: Input Parameters

2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.
3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. (See Figure 15.24).
4. Step 4 of 4: Click *Next* to annotate and save the entity list.(See Figure 15.25).

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on [Filter Probesets on Data Files](#)
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

15.3.3 Analysis

- **Statistical Analysis**
For details refer to section [Statistical Analysis](#) in the advanced workflow.
- **Filter on Volcano Plot**
For details refer to section [Filter on Volcano Plot](#)

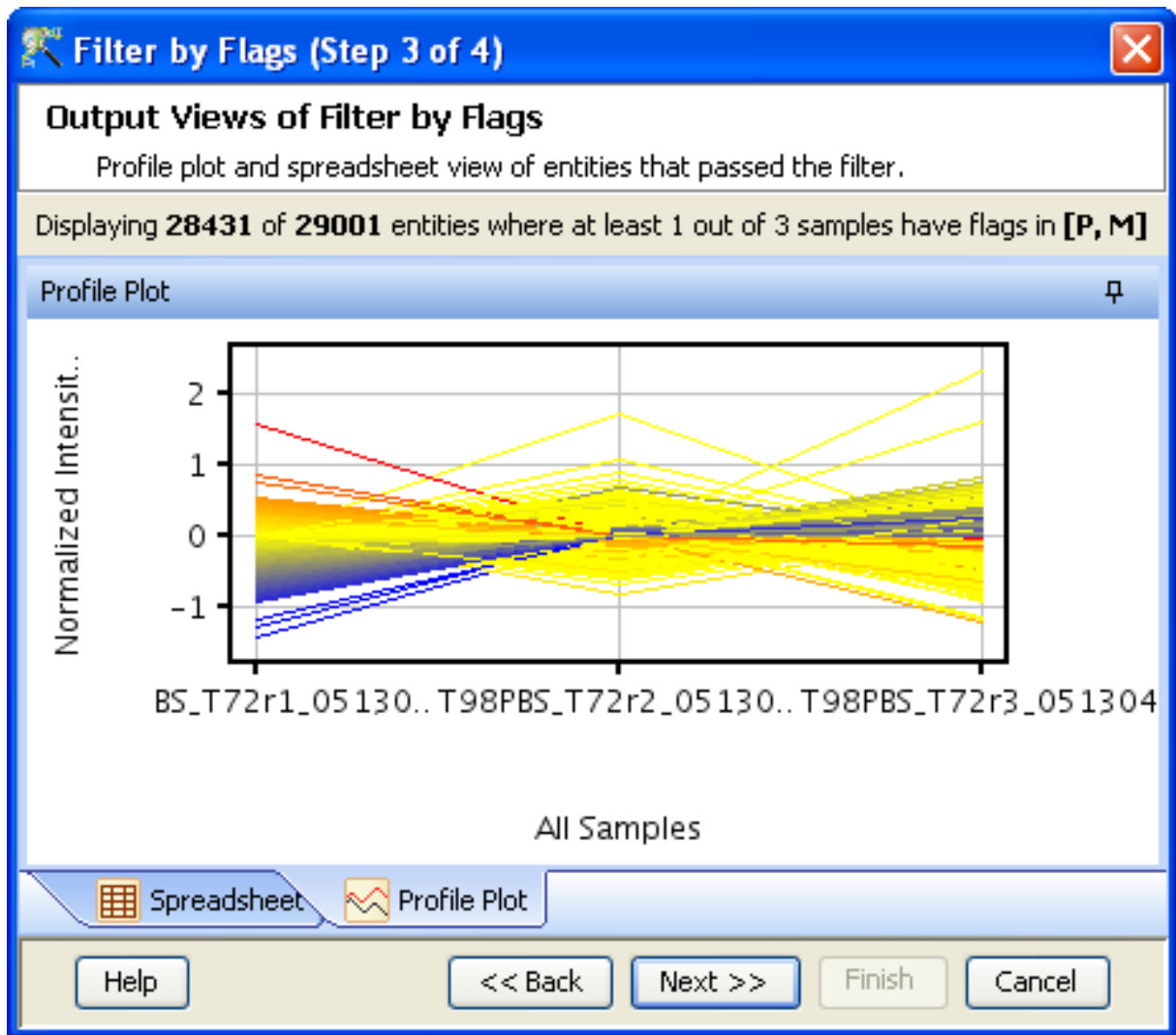


Figure 15.24: Output Views of Filter by Flags

- **Fold Change**
For details refer to section [Fold Change](#)
- **Clustering**
For details refer to section [Clustering](#)
- **Find Similar Entities**
For details refer to section [Find Similar Entities](#)
- **Filter on Parameters**
For details refer to section [Filter on Parameters](#)
- **Principal Component Analysis**
For details refer to section [PCA](#)

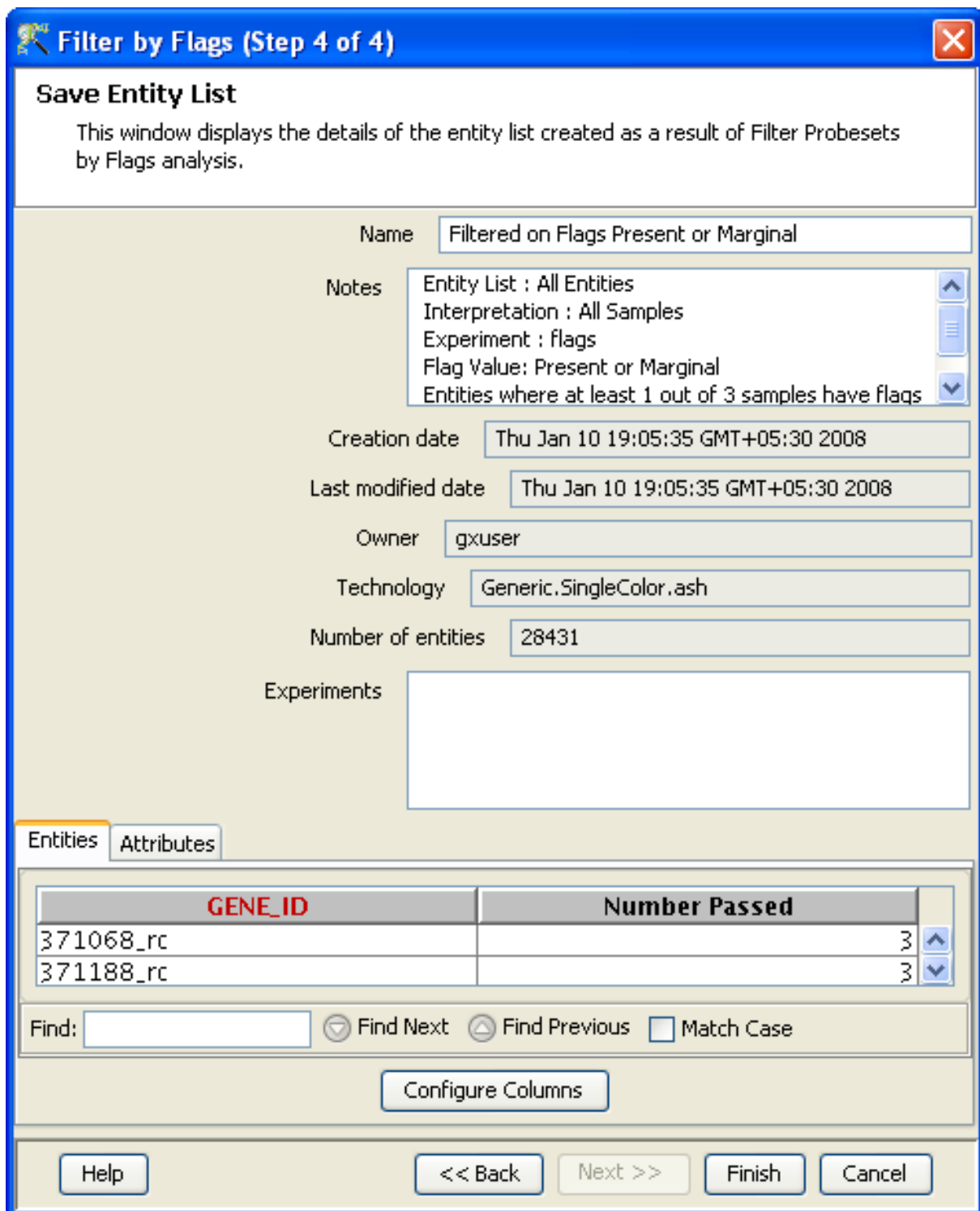


Figure 15.25: Save Entity List

15.3.4 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

15.3.5 Results

- **Gene Ontology (GO) analysis**
GO is discussed in a separate chapter called [Gene Ontology Analysis](#).
- **Gene Set Enrichment Analysis (GSEA)**
Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).
- **Gene Set Analysis (GSA)**
Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).
- **Pathway Analysis**
Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).
- **Find Similar Entity Lists**
This feature is discussed in a separate section called [Find Similar Entity Lists](#)
- **Find Significant Pathways**
This feature is discussed in a separate section called [Find Significant Pathways](#).
- **Launch IPA**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Import IPA Entity List**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Extract Interactions via NLP**
This feature is discussed in detail in the chapter [Pathway Analysis](#).

15.3.6 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)
- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)
- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)
- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

Chapter 16

Analyzing Generic Two Color Expression Data

GeneSpring GX supports Generic Two color experiments, such as spotted cDNA arrays. However, a technology first needs to be created, based upon the file format being imported.

16.1 Creating Technology

Technology creation is a step common to both Generic Single Color and Two color experiments. Technology creation enables the user to specify the columns (Signals, Flags, Annotations etc.) in the data file and their configurations which are to be imported. Different technologies need to be created for different file formats. Custom technology can be created by navigating to *Annotations* in the toolbar and selecting *Create Technology* → *Custom from file*. **GeneSpring GX** also allows the user to create a technology specifically for GPR files via *Annotations* → *Create Technology* → *From .gpr files*. This technology can later be used for creating a Generic Two Color experiment.

The process of creating a technology uses one data file as a sample file to mark the columns. Therefore, it is important that all the data files being used to create an experiment should have identical formats.

Technology creation using both the methods is detailed below:

16.1.1 Creation of Custom Technology-Non gpr files

The *Create Custom Technology* wizard has multiple steps. While steps 1, 2, 3 and 9 are common to both the Single color and Two Color, the remaining steps are specific to either of the two technologies.

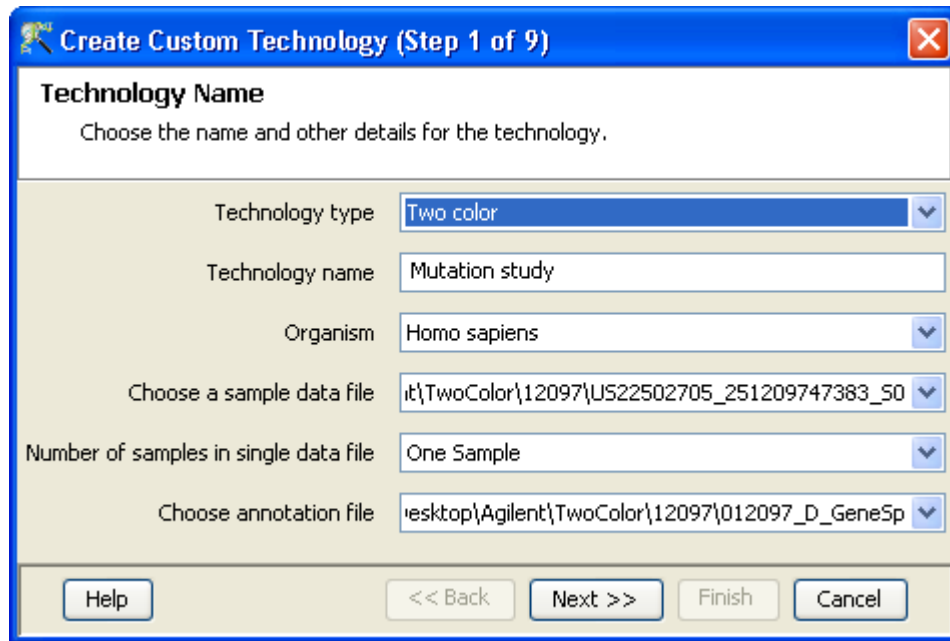


Figure 16.1: Technology Name

- **Technology Name (Step 1 of 9):** User input details, i.e., Technology type, Technology name, Organism, Sample data file location, Number of samples in a single data file and particulars of the annotation file are specified here. Click *Next*. See Figure 16.1
- **Format data set (Step 2 of 9):** This allows the user to specify the data file format. For this operation, four options are provided, namely, the *Separator*, the *Text qualifier*, the *Missing Value Indicator* and the *Comment Indicator*. The *Separator* option specifies if the fields in the file to be imported are separated by a tab, comma, hyphen, space etc. New separators can be defined by scrolling down to **Enter New** and providing the appropriate symbol in the textbox. *Text qualifier* is used for indicating characters used to delineate full text strings. This is typically a single or double quote character. The *Missing Value Indicator* is for declaring a string that is used whenever a value is missing. This applies only to cases where the value is represented explicitly by a symbol such as N/A or NA. The *Comment Indicator* specifies a symbol or string that indicates a comment section in the input file. Comment Indicators are markers at the beginning of the line which indicate that the line should be skipped (typical examples is the # symbol). See Figure 16.2
- **Select Row Scope for Import (Step 3 of 9):** The data files typically contains headers which are descriptive of the chip type and are not needed for the analysis. Only those rows containing the data values are required. The purpose of this step is to identify which rows need to be imported. The rows to be imported must be contiguous in the file. The rules defined for importing rows from this file will then apply to all other files to be imported using this technology. Three options are provided for selecting rows:

The default option is to select all rows in the file. Alternatively, one can choose to take a block of rows between specific row numbers (use the preview window to identify row numbers) by entering the row numbers in the appropriate textboxes. Remember to press the Enter key before proceeding. In addition, for situations where the data of interest lies between specific text markers, those text markers

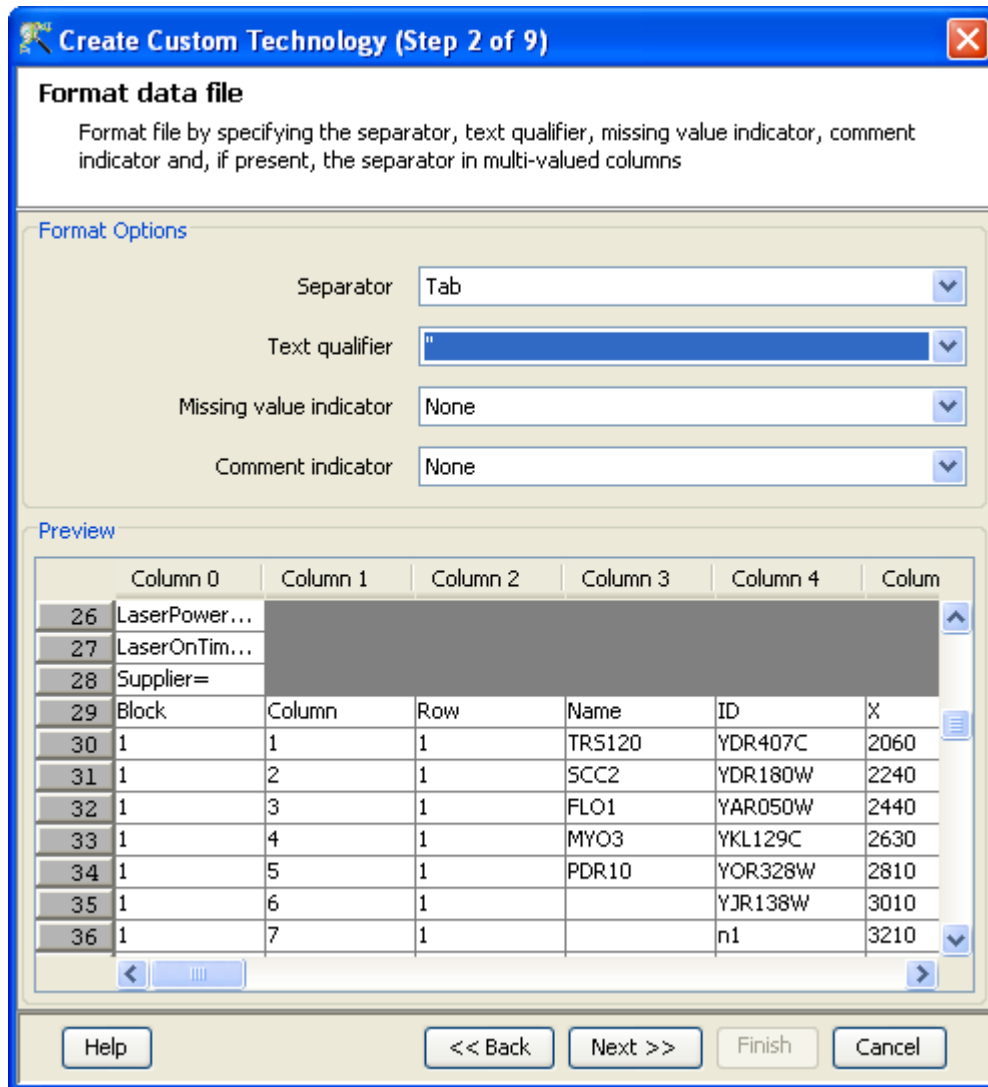


Figure 16.2: Format data file

can be indicated. Note also that instead of choosing one of the options from the radio buttons, one can choose to select specific contiguous rows from the preview window itself by using Left-Click and Shift-Left-Click on the row header. The preview shows only the first 100 rows of the file by default. The user can change the default settings from *Tools* → *Options* → *Miscellaneous* → *Custom Data Library Creation* → *Number of preview lines*. The panel at the bottom should be used to indicate whether or not there is a header row; in the latter case, dummy column names will be assigned. See Figure 16.3.

- Steps 4 and 5 are used while creating custom technology for a single color experiment.
- **Create Custom technology (Step 6 of 9):** After the rows to be imported have been identified, columns for the gene identifier, background (BG) corrected signals and flag values for Cy5 and Cy3 channels in the data file have to be indicated. In case of a file containing a single flag column either the flag Cy3 or flag Cy5 can be used to mark the same. Categories within the flag columns can

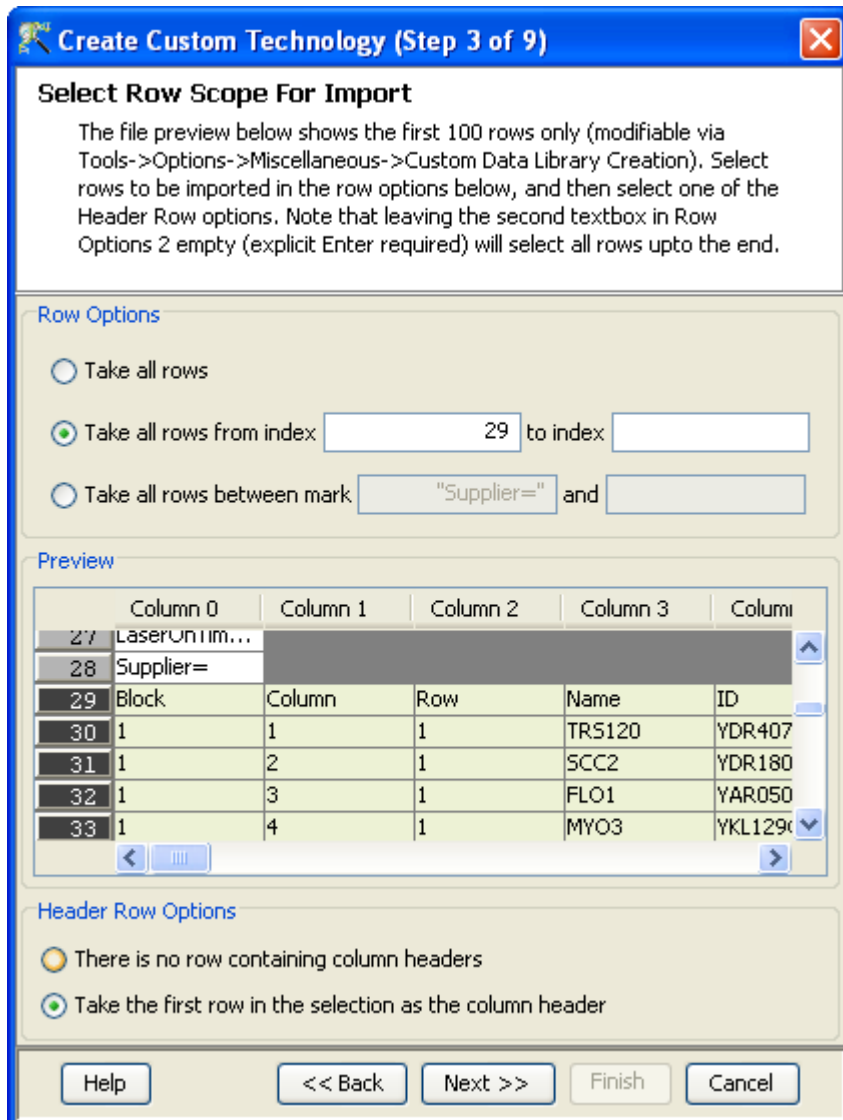


Figure 16.3: Select Row Scope for Import

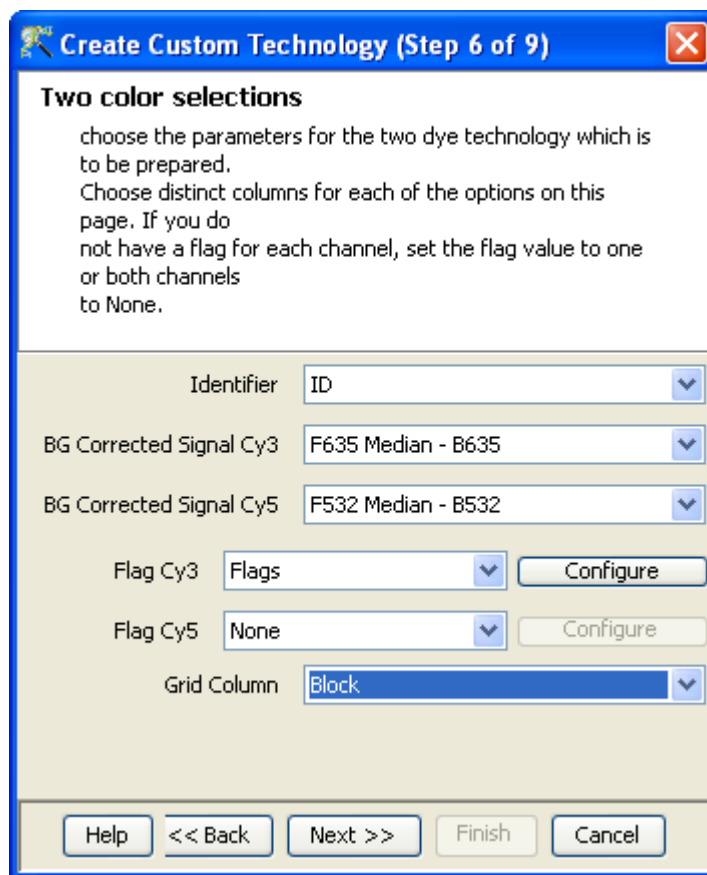


Figure 16.4: Two Color Selections

be configured to designate Present (P), Absent (A) or Marginal (M) values. Grid column can be specified to enable block by block normalization. See Figure 16.4

Lowess sub-grid normalization can be performed by choosing the grid column.

Annotation column options have to be specified from steps 7 to 9.

- **(Step 7 and 8 of 9):** These steps are similar to the step 2 of 9 and is used to format the annotation file. If a separate annotation file does not exist, then the same data file can be used as an annotation file, provided it has the annotation columns.
- **(Step 8 of 9):** Identical to step 3 of 9, this allows the user to select row scope for import in the annotation file.
- **(Step 9 of 9):** The Step 9 of technology creation is an extremely important step which allows the user to mark the columns appropriately. Proper marking of the various columns will enable the various functionalities like GO, GSEA, Genome Browser, Pathway Analysis to proceed smoothly. The markings to be given for all these functions is elaborated below:

- GSEA: The annotation file should contain a column containing Gene Symbol. This column should be marked as Gene Symbol from the drop-down menu.
- GSA: The annotation file should contain a column containing the gene Symbol. This column should be marked as Gene Symbol from the drop-down menu.
- GO: For carrying out GO analysis, the annotation file can either contain a single column with all the GO IDs in it, separated by a separator or it can contain separate columns for the different GO processes. The single column with multiple GO IDs should be marked as *Gene Ontology accession* from the drop-down menu. Instead if columns containing individual GO processes(Biological Process, Cellular Component and Molecular Function) are present, they should be marked accordingly in the dropdown menu.
- Genome Browser: In order to view the data in Genome Browser, the annotation file should contain a Chromosome Start Index, Chromosome End Index, Strand and Chromosome Number columns. Provide the column mark for Chromosome Start index, Chromosome End index, Strand, Chromosome number respectively, from the drop-down menu.

Note: The Chromosome Start Index < Chromosome End Index. For viewing *Profile* track only, in the Genome Browser, chromosome start index and chromosome number are needed. The labelling of the chromosome numbers should follow this convention-chr1, chr2i.e. the word starts with chr followed by the chromosome number (without any space). For viewing data track, all four Chromosome Start Index, Chromosome End Index, Strand, Chromosome Number are needed.

- If a custom technology is being created using an **Illumina** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

For viewing using the Genome Browser, the annotation files has three columns which have values for all four (Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand) Therefore before creating the custom experiment the user needs to parse these columns and create three new columns as follows :

Probe_Chr_Orientation– This column can be taken as it is. It should be marked as Strand.

Chromosome – A new column must be created wherein a 'chr' should be appended to each entry in the Chromosome column and this new column should be marked as Chromosome Number.

Probe_Coordinates– This column has each entry in the format a-b where a < b. Two new columns need to be created. one which has only the a values, (it should be marked as Chromosome Start Index) one which has only the b values (it should be marked as Chromosome End Index).

- If a custom technology is being created using an **Agilent** data and annotation file, then for the Genome Browser functionality, the column markings have to be handled as follows:

The annotation files have a single column 'Map' which has values for all four Chromosome Start Index, Chromosome End Index and Chromosome Number and Strand. Therefore before creating the custom experiment the user needs to parse the file and separate the four columns as Chromosome Start Index, Chromosome End Index Chromosome Number and Strand.

Each entry in the Map column is typically in the format chrQ:a..b

if a < b, the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is a; the corresponding Chromosome End Index is b; the corresponding Strand is +.

if a > b the corresponding Chromosome Number is chrQ; the corresponding Chromosome Start Index is b; the corresponding Chromosome End Index is a; the corresponding Strand is - .

For example, a Map value of chr14:34101457..34101398 corresponds to a Chromosome Start Index of 34101398, a Chromosome End Index of 34101457, a Chromosome Number of chr14 and a Strand of - (because in chrX:a..b a>b)

For example, a Map value of chr6:46222041..46222100 corresponds to a Chromosome Start Index of 46222041, a Chromosome End Index of 46222100, a Chromosome Number of chr6 and a Strand of +(because in chrX:a..b a<b)

- **Import BioPAX pathways:** Pathways being imported should be in .owl format. During custom technology creation, provide the column mark for Entrez Gene ID/SwissProt from the drop-down menu. Only after this mark is provided can the proteins involved in a particular pathway be highlighted.
- **Find Significant Pathways:** The annotation file should contain an Entrez Gene ID/SwissProt column, which have to be marked appropriately as Entrez Gene ID/SwissProt.
- **Translation:** This operation can be performed between organisms listed in the **Homologene table** in section [Translation](#). Entrez Gene ID column has to be marked for performing translation.

See figure [16.6](#).

The types of Data and Attribute marks available for the annotation columns are

- **Categorical:** A column marked as a "categorical" column means that the values in this column fall into certain finite distinct categories.
- **Continuous:** A column marked as a "continuous" column means that the values in this column can vary, potentially, over any large range.
- **String:** A continuous sequence of symbols or digits, not including a space.
- **Float:** A real number, i.e a number which can be given by a decimal representation.

The annotation marks are colored on the basis of their functionality in the tool. The meaning of the various colors are provided in the figure [16.5](#). This figure is provided solely for visualization purposes and is not available from the tool.

Click ***Finish*** to exit the wizard.

16.1.2 GenePix Result Technology creation

This option allows the user to create a technology for files (.gpr) that have been generated using the GenePix Pro software. This feature is compatible for different versions of the gpr file. The gpr file used to create the technology should contain the following columns - ID, F635 Median - B635, F532 Median - B532, and Flags. This technology creation option is accessible from ***Annotations***→***Create Technology***→***From .gpr file***. On selecting this option, the user has to go through the following step for custom technology creation:

Mark Category Colors










Identifier	 192, 0, 0	▼
Signal	 0, 0, 192	▼
Co-ordinate	 0, 96, 0	▼
Quality	 0, 0, 96	▼
Exon Annotation	 0, 192, 0	▼
Gene Annotation	 128, 0, 128	▼
NetAffx	 128, 0, 128	▼
Label	 128, 64, 64	▼
Other	 0, 0, 0	▼

Figure 16.5: Annotation Mark Colors

- **Input Data(Step 1 of 1):**

This step allows the user to input data required for technology creation. The user has to provide the technology name, organism and the sample data file. See Figure 16.7. The organism name is optional, but in the event of Biological Genome creation, the organism name is a must. Biological Genome contains most of the annotations using which additional analysis like GO, GSEA etc can be performed. For further details, refer to the section on [Biological Genome](#).

The technology created does not have any annotations associated with it. The user can add annotations via *Annotations*→*Update Technology Annotations*→*From file or Biological Genome*. For more details on the same, refer to [Update Technology Annotations](#)

16.1.3 Project and Experiment Creation

After technology creation, data files satisfying the file format can be used to create an experiment. The following steps will guide you through the process of experiment creation.

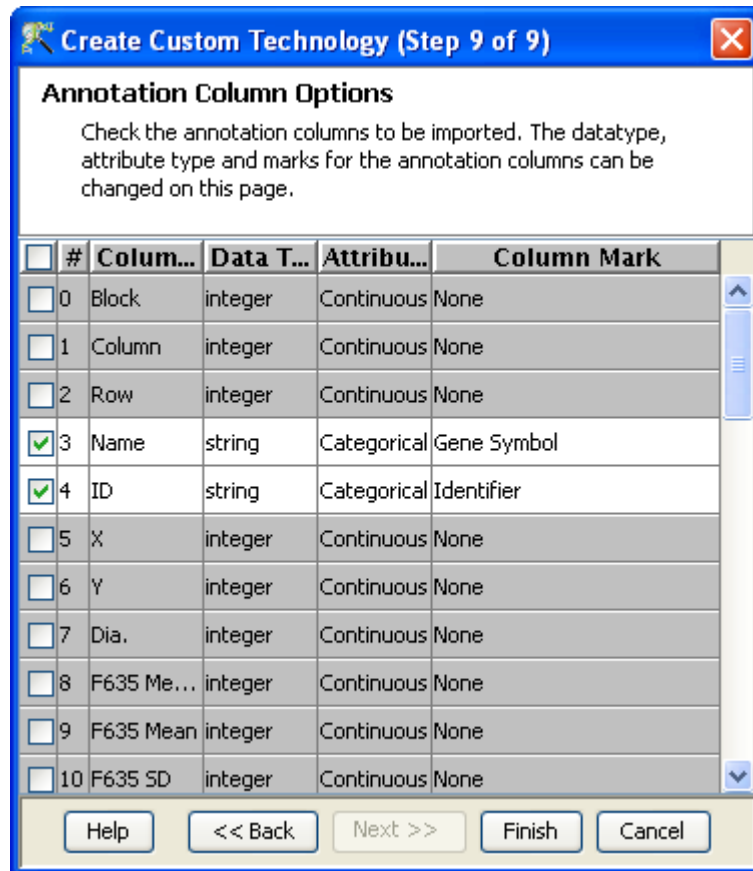


Figure 16.6: Annotation Column Options

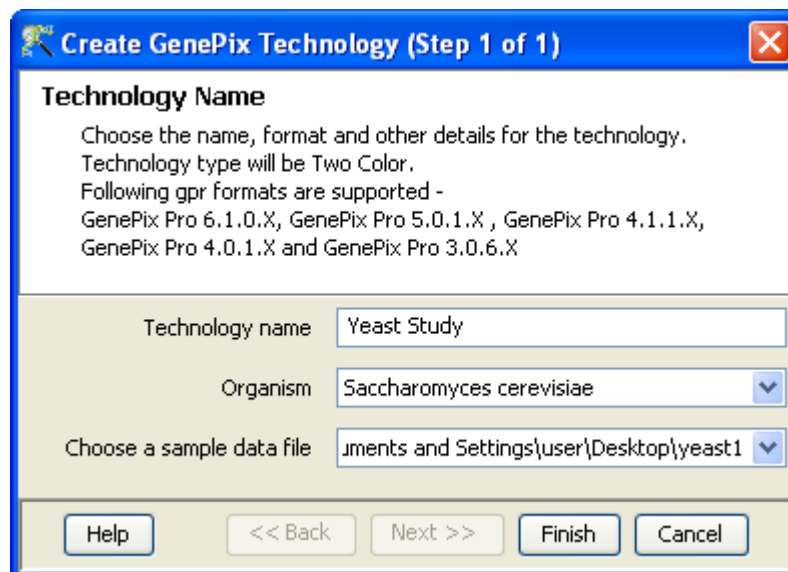


Figure 16.7: Technology Creation



Figure 16.8: Welcome Screen

Upon launching **GeneSpring GX** , the startup is displayed with 3 options. See Figure 16.8

1. **Create new project**
2. **Open existing project**
3. **Open recent project**

Either a new project can be created or else a previously generated project can be opened and re-analyzed. On selecting *Create New Project*, a window appears in which details (name of the project and notes) can be recorded. Press *OK* to proceed. See Figure 16.9

An Experiment Selection Dialog window then appears with two options

1. **Create new experiment**
2. **Open existing experiment**

See Figure 16.10

Selecting *Create new experiment* allows the user to create a new experiment (steps described below). *Open existing experiment* allows the user to use existing experiments from any previous projects in the current project. Choosing *Create new experiment* opens up a New Experiment dialog in which *Experiment*

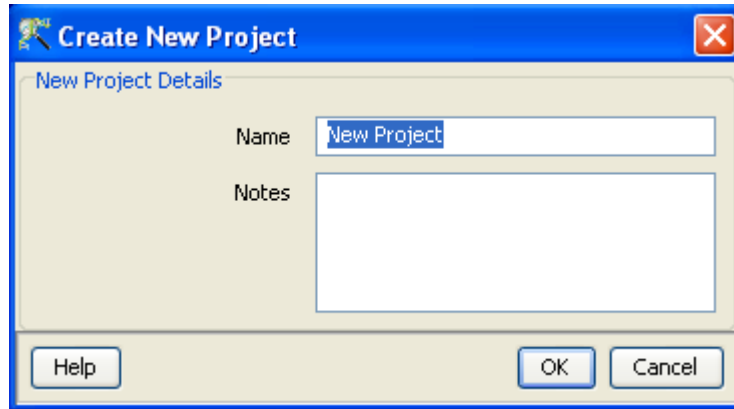


Figure 16.9: Create New project

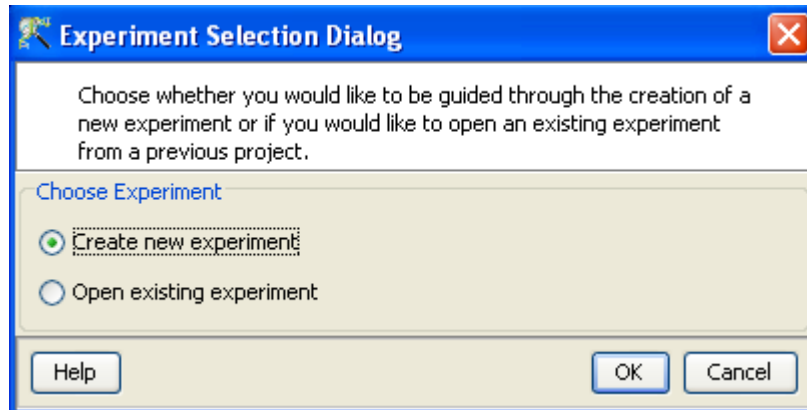


Figure 16.10: Experiment Selection

name can be assigned. The *Experiment type* should then be specified (Generic Two Color), using the drop down button. The *Workflow Type* that appears is the *Advanced* type. Unlike the other technologies where *Guided* and *Advanced* analysis workflows are available, in case of Generic Two-color, only the *Advanced Workflow* is supported . Click *OK* will open a new experiment wizard. See Figure 16.11

16.2 Advanced Analysis

The *Advanced Workflow* offers a variety of choices to the user for the analysis. Thresholding can be performed. Based upon the technology, Lowess or sub-grid Lowess normalization can be performed. Additionally there are options for baseline transformation of the data and for creating different interpretations.

The *New Experiment Wizard* has the following steps:

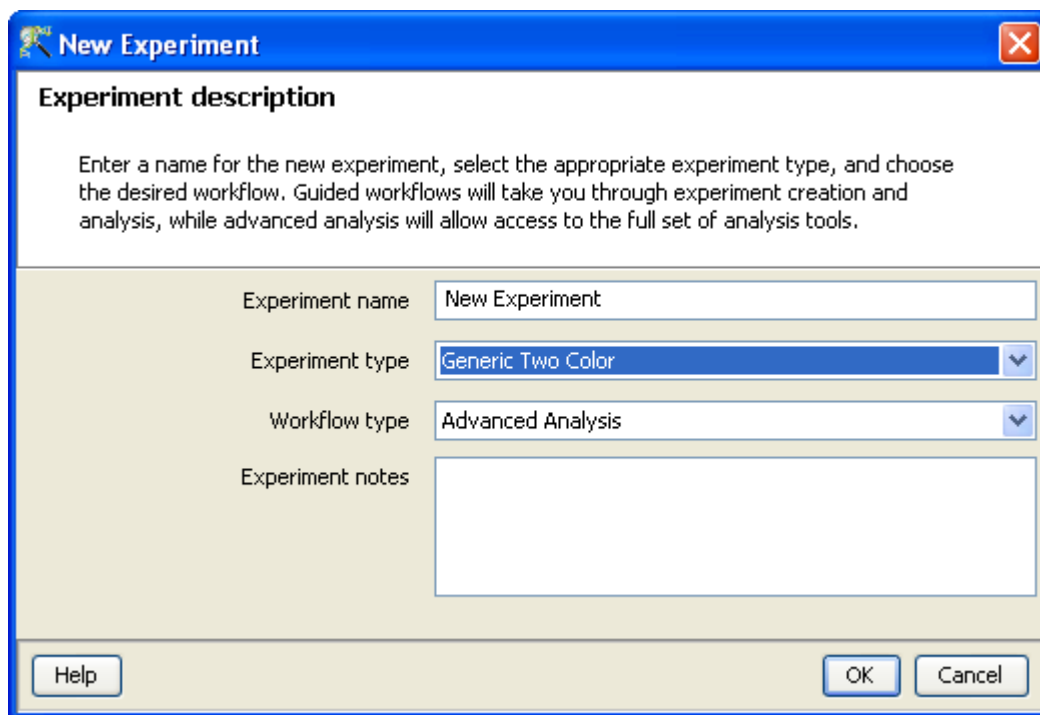


Figure 16.11: Experiment Description

1. **New Experiment (Step 1 of 4):** The technology (created as mentioned above) can be selected and the new data files or previously used data files in **GeneSpring GX** can be imported in to create the experiment. A window appears containing the following options:
 - (a) **Choose Files(s)**
 - (b) **Choose Samples**
 - (c) **Choose Raw Files**
 - (d) **Reorder**
 - (e) **Remove**

An experiment can be created using either the data files or else using samples. Upon loading data files, **GeneSpring GX** associates the files with the technology (see below) and creates samples. These samples are stored in the system and can be used to create another experiment via the *Choose Samples* option through a sample search wizard. If the user has imported any custom experiments from **GeneSpring GX 7** and wants to recreate the experiment in **GeneSpring GX**, then the user can create a new technology in the tool with an original raw file and later utilize the **Choose Raw Files** option to choose the raw files associated with the migrated custom experiment. For selecting data files and creating an experiment, click on the *Choose File(s)* button, navigate to the appropriate folder and select the files of interest. Select *OK* to proceed.

The sample search wizard that comes up via the option *Choose Samples* has the following search conditions:

- (a) **Search field** (which searches using any of the 6 following parameters- (Creation date, Modified date, Name, Owner, Technology, Type).

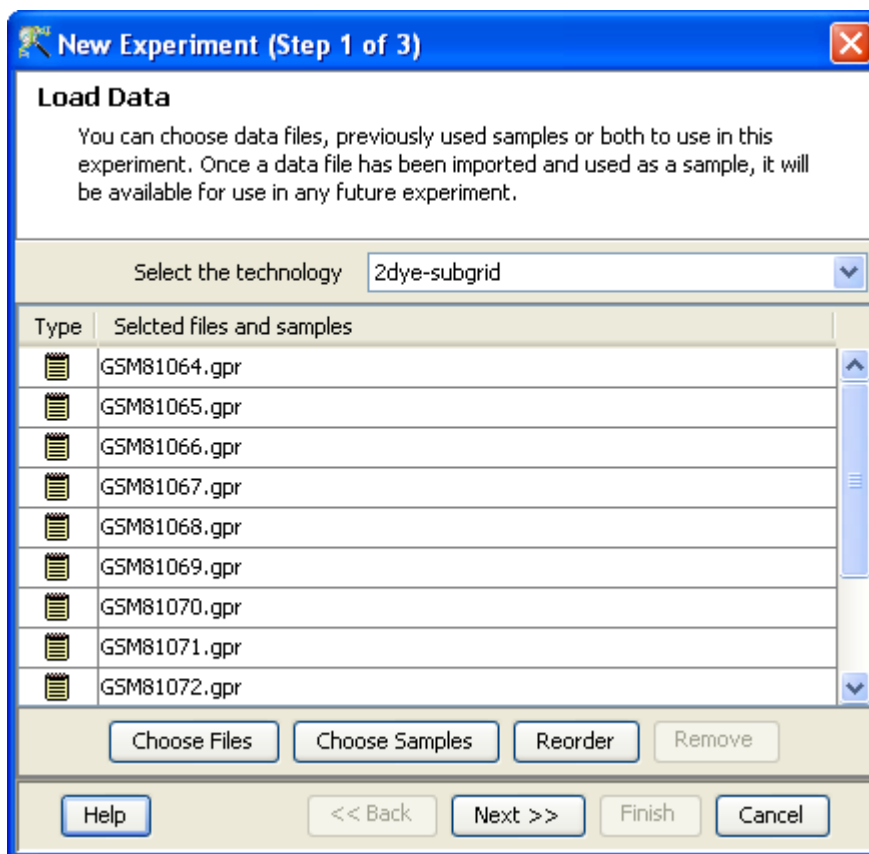


Figure 16.12: Load Data

- (b) **Condition** (which requires any of the 4 parameters-Equals, Starts with, Ends with and includes Search value).
- (c) **Value**

Multiple search queries can be executed and combined using either *AND* or *OR*.

Samples obtained from the search wizard can be selected and added to the experiment using *Add* button, similarly can be removed using *Remove* button.

After selecting the files, clicking on the *Reorder* button opens a window in which the particular sample or file can be selected and can be moved either up or down by pressing on the buttons. Click on *OK* to enable the reordering or on *Cancel* to revert to the old order. See Figure 16.12

2. **New experiment (Step 2 of 4):** Dye swap arrays, if any, can be indicated in this step. Data/Sample files chosen in previous step are shown here and the user can select those arrays that were dye-swapped while performing the experiment. Accordingly, **GeneSpring GX** will swap the data between cy5 and cy3 for these arrays. See Figure 16.13
3. **New experiment (Step 3 of 4):** This gives the options for preprocessing of input data. It allows the user to threshold raw signals to chosen values and the selection of Lowess normalization. In case of experiment creation using .gpr files, the option to perform sub-grid Lowess is not present.

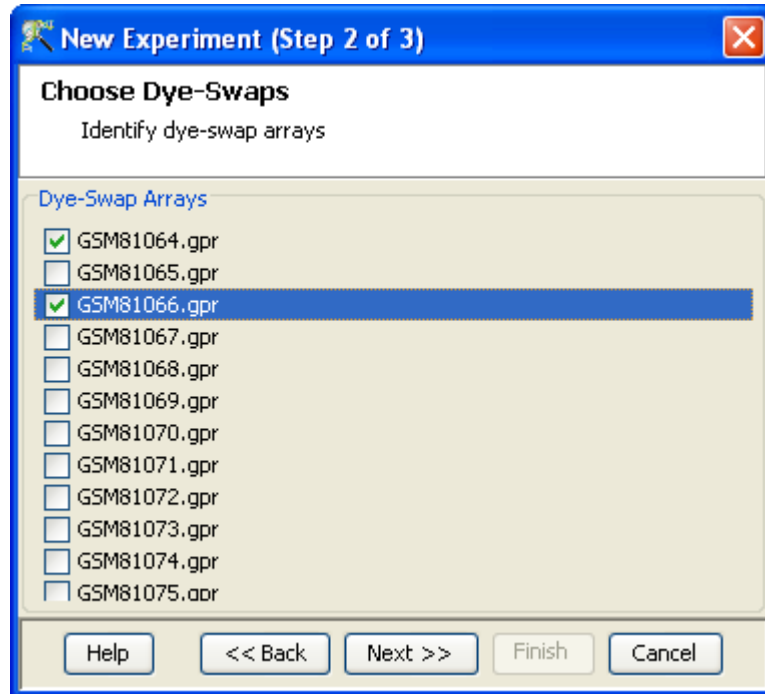


Figure 16.13: Choose Dye-Swaps

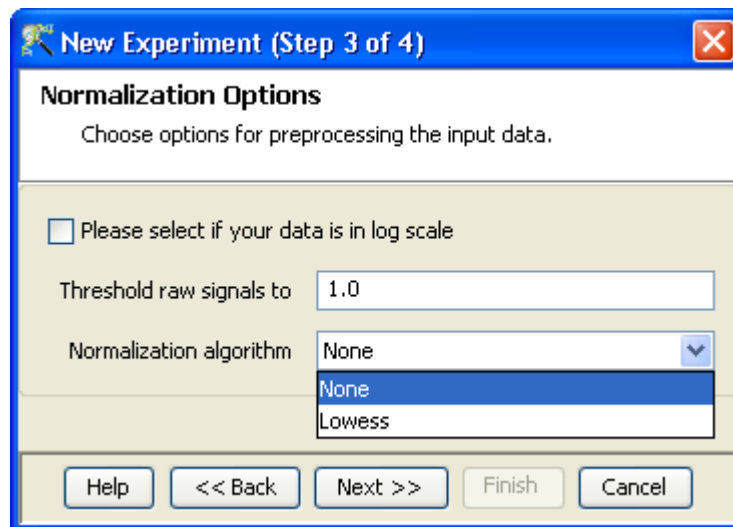


Figure 16.14: Preprocess Options

In case, the data is already log transformed, the user can select the checkbox stating that their signal values are already in log scale. This will disable the thresholding option also.

See Figure 16.14

4. New experiment (Step 4 of 4):

This step provides the baseline options which include:

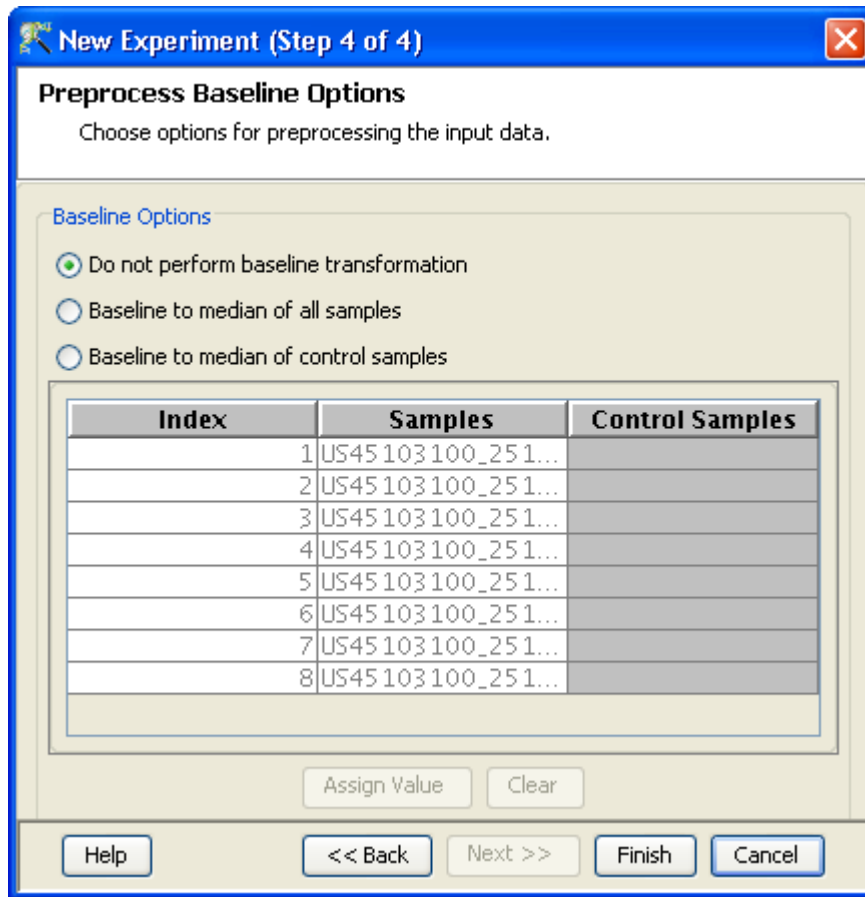


Figure 16.15: Preprocess Baseline Options

- *Do not perform baseline*
- *Baseline to median of all samples:* For each probe the median of the log summarized values from all the samples is calculated and subtracted from each of the samples.
- *Baseline to median of control samples:* For each sample, an individual control or a set of controls can be assigned. Alternatively, a set of samples designated as controls can be used for all samples. For specifying the control for a sample, select the sample and click on **Assign value**. This opens up the **Choose Control Samples** window. The samples designated as Controls should be moved from the *Available Items* box to the *Selected Items* box. Click on **Ok**. This will show the control samples for each of the samples.

In *Baseline to median of control samples*, for each probe the median of the log summarized values from the control samples is first computed and then this is subtracted from the sample. If a single sample is chosen as the control sample, then the probe values of the control sample are subtracted from its corresponding sample.

See Figure 16.15

Clicking *Finish* creates an experiment, which is displayed as a Box Whisker plot in the active view. Alternative views can be chosen for display by navigating to *View* in Toolbar.

16.2.1 Data Processing for Generic Two Color Data

1. **File formats:** The files should be tabular in nature. For example, .csv, .tsv , .gpr etc.
2. **Signal Columns:** When custom technology is created via *Annotations*→*Create Technology* →*From .gpr file*, the signal columns taken from files are F532 Median-B532 for cy3 and F635 Median-B635 for cy5.
3. **Raw:** The term "raw" signal values refer to the linear data after thresholding to 1.0 and summarization for the individual channels (cy3 and cy5). Summarization is performed by computing the geometric mean.
4. **Normalized:** The term Normalized signal value refers to the raw data after normalization of cy5 channel, ratio computation (cy5/cy3), log transformation and Baseline Transformation.
5. **Treatment of on-chip replicates:** The signal value of a probeset is the geometric mean of all its probes.
6. **Flag values:** The values for the probes are configured by the user during the creation of technology as either present, marginal or absent. Based on the values of the probes, the probeset is assigned a flag value. The order of importance for flag values for probes in a probeset is *Present*>*Marginal*>*Absent*. When custom technology is created via *Annotations*→*Create Technology* →*From .gpr file*, flags are configured by the tool. A value of -50 is designated as Marginal(M) and anything below is considered Absent(A) and anything above is considered as Present(P).
7. **Treatment of Control probes:** The control probes are included while performing normalization.
8. **Empty Cells:** Empty cells might be present in the intensity values column for certain genes in the data file. These genes are brought in **GeneSpring GX** . But an entity list containing these genes cannot be used for running clustering and class prediction.
9. **Sequence of events:** The sequence of events involved in the processing of the data files is: thresholding→summarization→normalization→ratio computation→log transformation→Baseline Transformation.
10. **Merging of files:** Multiple files in Generic experiment creation are combined based on the Identifier column using the following rules. The very first file among the various files chosen server as a master reference (you can determine which file serves as the first file using the *Reorder* button on Page 1 of the New Experiment Creation page). The number of rows in this master must exceed the number of rows in all subsequent files, for extra rows in these subsequent files are dropped. Next, all identifiers in the Identifier column of this first file are considered and missing values in these, if any, are discarded. This results in a set of valid identifier values; all rows in all other files whose identifier values are outside of this set are discarded. Next, on-chip replicates are determined by counting the number of occurrences of each valid identifier in the first file. Consider for example an identifier Id1 which appears 3 times in file 1. Then rows corresponding to the first 3 occurrences of Id1 are taken in each of the other files; if there are fewer than 3 rows, then as many rows that are present are taken; and if there are more than 3 rows, then the first 3 are taken. The summarized value for Id1 in each file is determined by taking a geometric mean over these chosen rows.

16.2.2 Experiment Setup

- **Quick Start guide:** Clicking on this link will take you to the appropriate chapter in the on-line manual giving details of loading expression files into **GeneSpring GX**, the Advanced workflow, the method of analysis, the details of the algorithms used and the interpretation of results
- **Experiment Grouping:** *Experiment parameters* defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#)
- **Create Interpretation:** An interpretation specifies how the samples would be grouped into experimental conditions for display and used for analysis. For details refer to the section on [Create Interpretation](#)
- **Create New Gene Level Experiment:** Allows creating a new experiment at gene level using the probe level data in the current experiment.

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.
 - * Chromosome Start Index
 - * Chromosome End Index
 - * Chromosome Map
 - * Cytoband
 - * Probe Sequence
- Flag information will also be dropped.
- Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
- Experiment grouping, if present in the original experiment, will be retained.
- The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

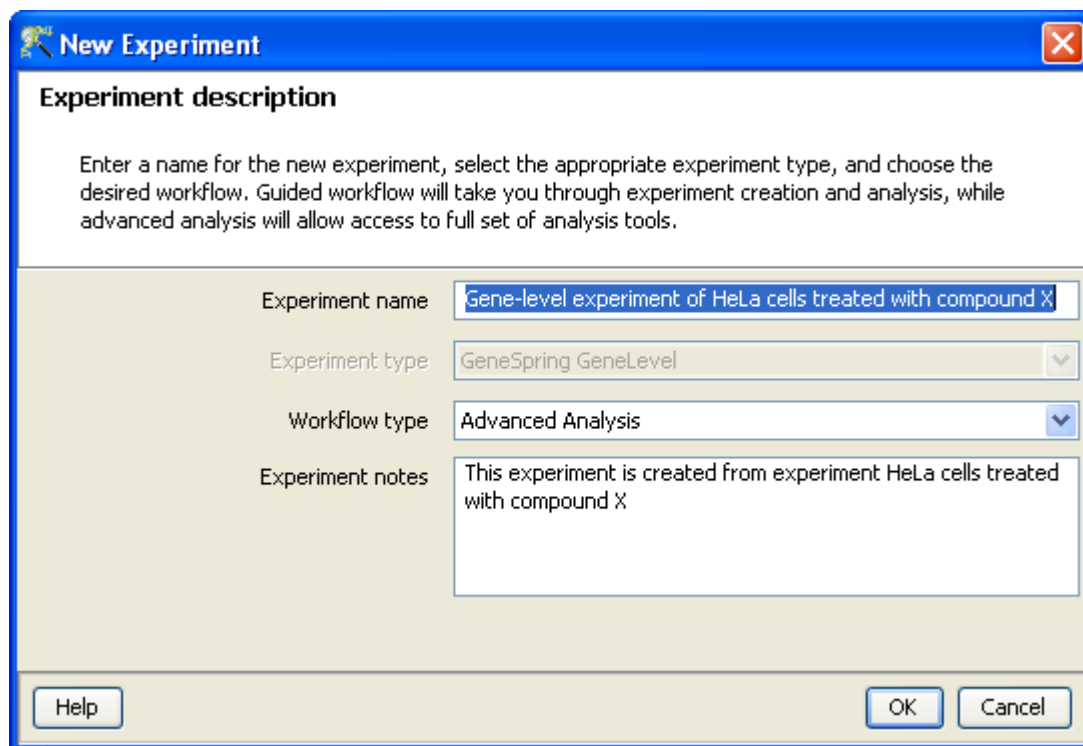


Figure 16.16: Gene Level Experiment Creation

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.

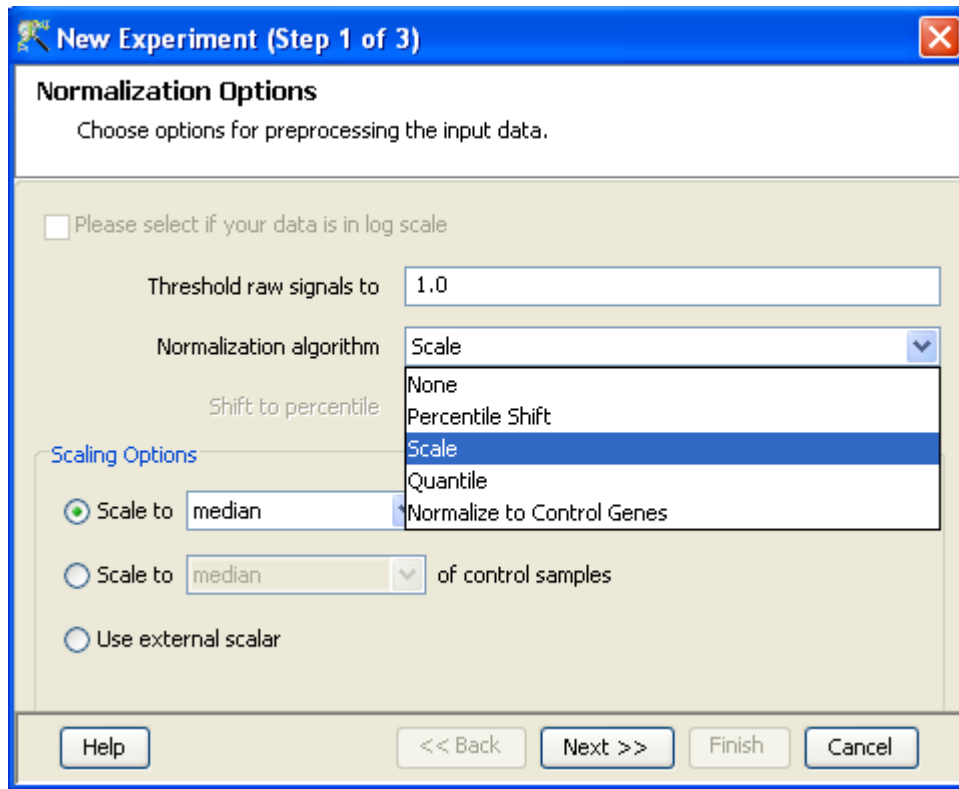


Figure 16.17: Gene Level Experiment Creation - Normalization Options

- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations.

Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

For two colour, raw values are summarized for each channel separately and then log ratios are taken.

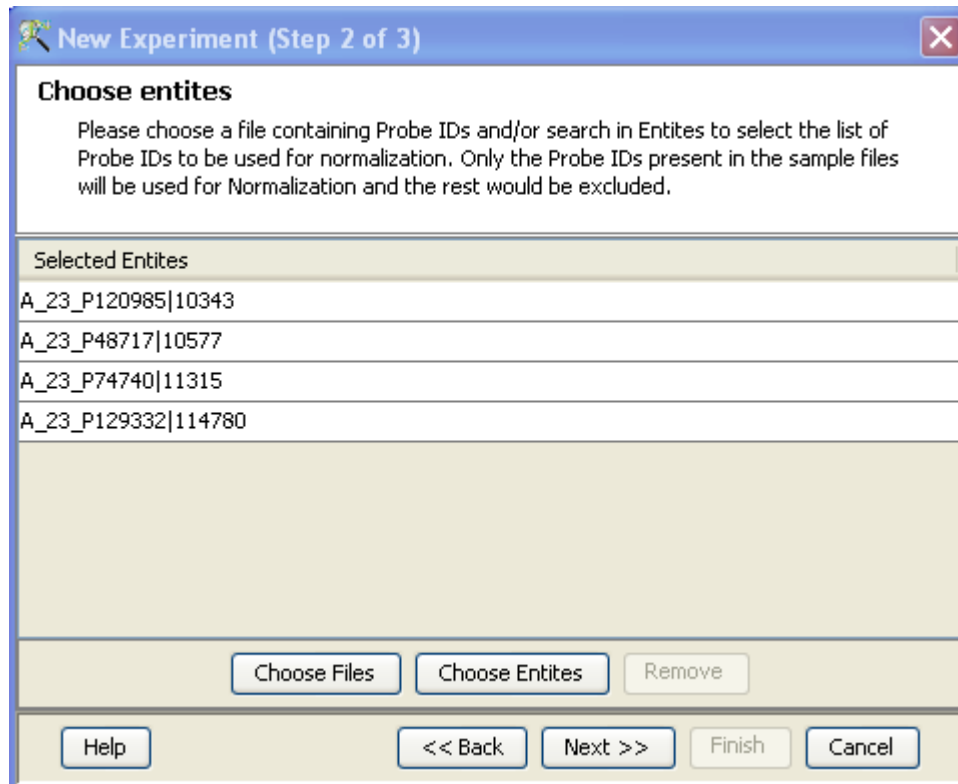


Figure 16.18: Gene Level Experiment Creation - Choose Entities

16.2.3 Quality Control

- **Quality Control on Samples:**

The view shows four tiled windows:

1. Experiment grouping
2. PCA scores
3. Legend

See Figure [16.20](#)

Experiment Grouping shows the parameters and parameter values for each sample.

Principal Component Analysis (PCA) calculates the PCA scores and visually represents them in a 3D scatter plot. The scores are used to check data quality. It shows one point per array and is colored by the *Experiment Factors* provided earlier in the *Experiment Groupings* view. This allows viewing of separations between groups of replicates. Ideally, replicates within a group should cluster together and separately from arrays in other groups. The PCA components, represented in the X, Y and Z axes are numbered 1, 2, 3... according to their decreasing significance. The 3D PCA scores plot can be customized via **Right-Click**→**Properties**. To zoom into a 3D Scatter plot, press the Shift key and simultaneously hold down the left mouse button and move the mouse upwards. To

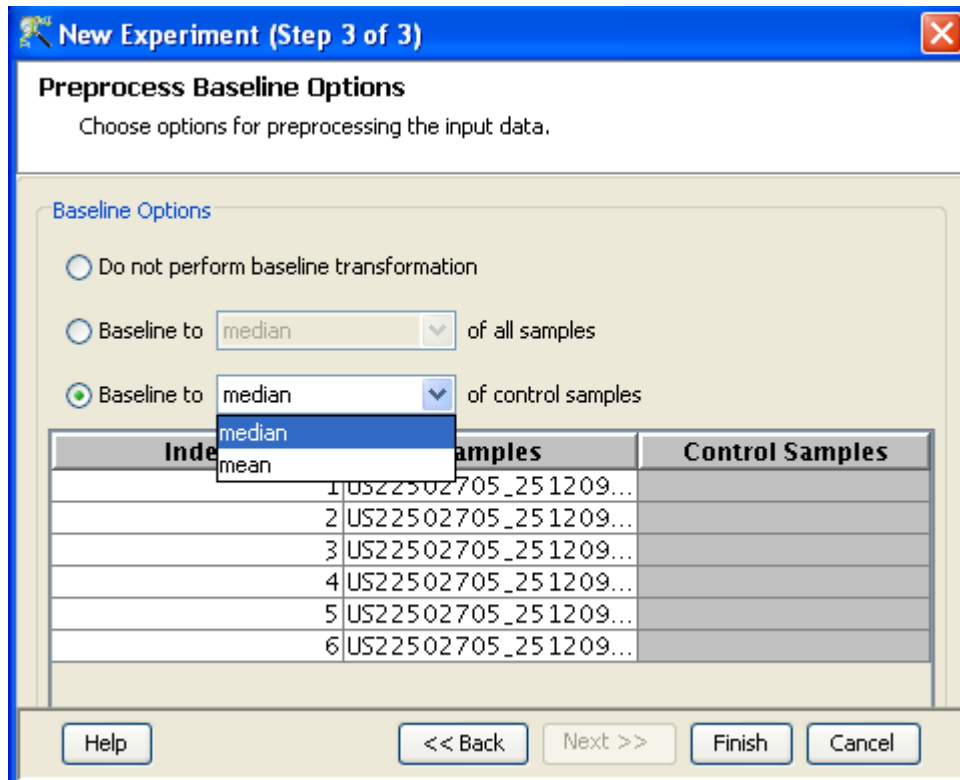


Figure 16.19: Gene Level Experiment Creation - Preprocess Baseline Options

zoom out, move the mouse downwards instead. To rotate, press the Ctrl key, simultaneously hold down the left mouse button and move the mouse around the plot.

The fourth window shows the legend of the active QC tab.

Click on **OK** to proceed.

- **Filter Probe Set by Expression:** Entities are filtered based on their signal intensity values. for details refer to the section on [Filter Probesets by Expression](#)
- **Filter Probe Set by Flags:**

In this step, the entities are filtered based on their flag values, the P(present), M(marginal) and A(absent). Users can set what proportion of conditions must meet a certain threshold. The flag values that are defined at the creation of the new technology (Step 2 of 3) are taken into consideration while filtering the entities. The filtration is done in 4 steps:

1. Step 1 of 4 : *Entity list and interpretation* window opens up. Select an entity list by clicking on *Choose Entity List* button. Likewise by clicking on *Choose Interpretation* button, select the required interpretation from the navigator window. This is seen in Figure 16.21
2. Step 2 of 4: This step is used to set the Filtering criteria and the stringency of the filter. Select the flag values that an entity must satisfy to pass the filter. By default, the Present and Marginal flags are selected. Stringency of the filter can be set in *Retain Entities* box.(See Figure 16.22) .
3. Step 3 of 4: A spreadsheet and a profile plot appear as 2 tabs, displaying those probes which have passed the filter conditions. Baseline transformed data is shown here. Total number of probes

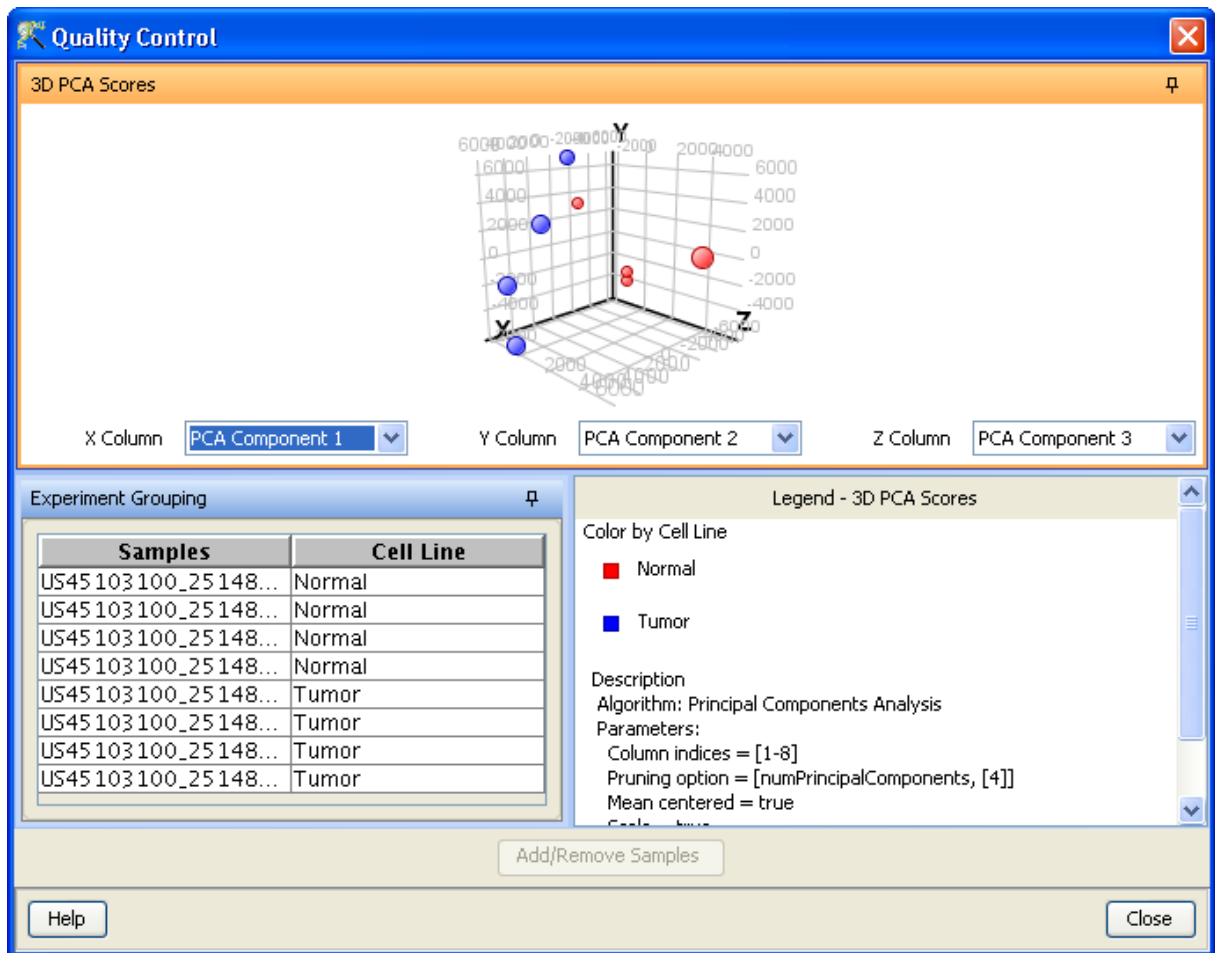


Figure 16.20: Quality Control

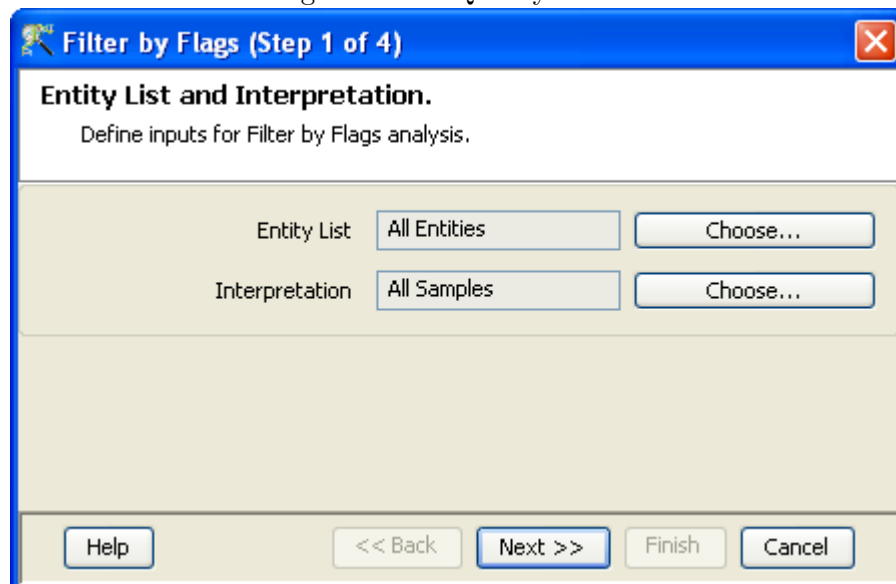


Figure 16.21: Entity list and Interpretation

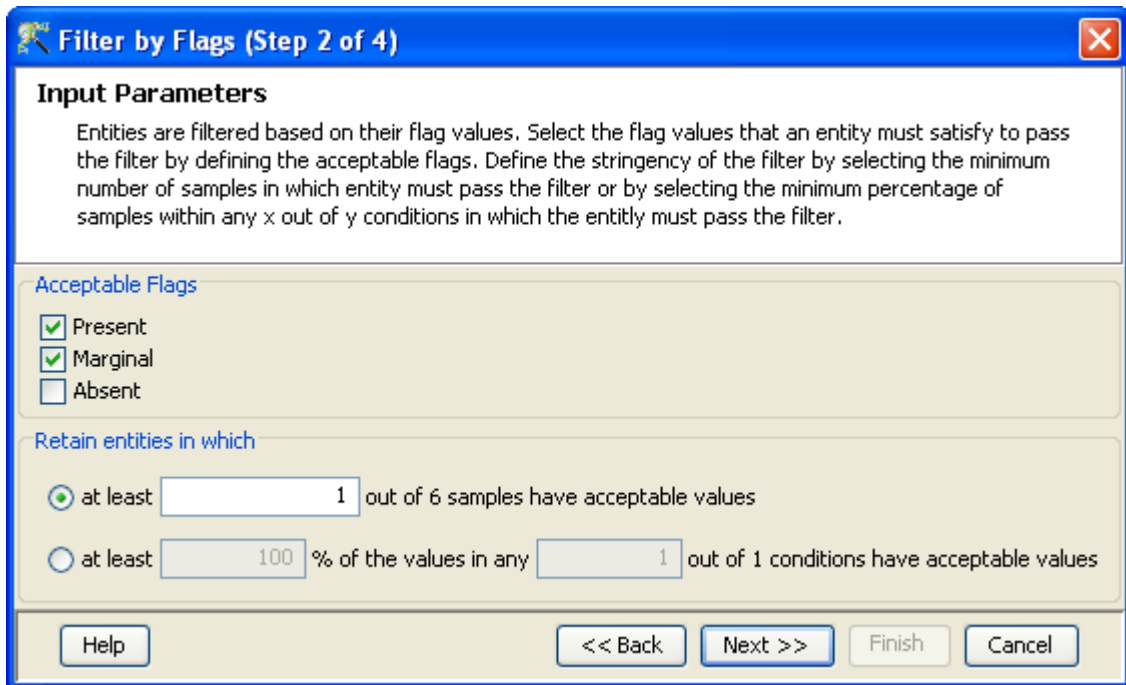


Figure 16.22: Input Parameters

and number of probes passing the filter are displayed on the top of the navigator window. (See Figure 16.23).

4. Step 4 of 4: Click *Next* to annotate and save the entity list. (See Figure 16.24).

- **Filter Probesets on Data Files:** Entities can be filtered based on values in a specific column of the original data files. For details refer to the section on [Filter Probesets on Data Files](#)
- **Filter Probesets by Error:** Entities can be filtered based on the standard deviation or coefficient of variation using this option. For details refer to the section on [Filter Probesets by Error](#)

16.2.4 Analysis

- **Statistical Analysis**

For details refer to section [Statistical Analysis](#) in the advanced workflow.

- **Filter on Volcano Plot**

For details refer to section [Filter on Volcano Plot](#)

- **Fold Change**

For details refer to section [Fold Change](#)

- **Clustering**

For details refer to section [Clustering](#)

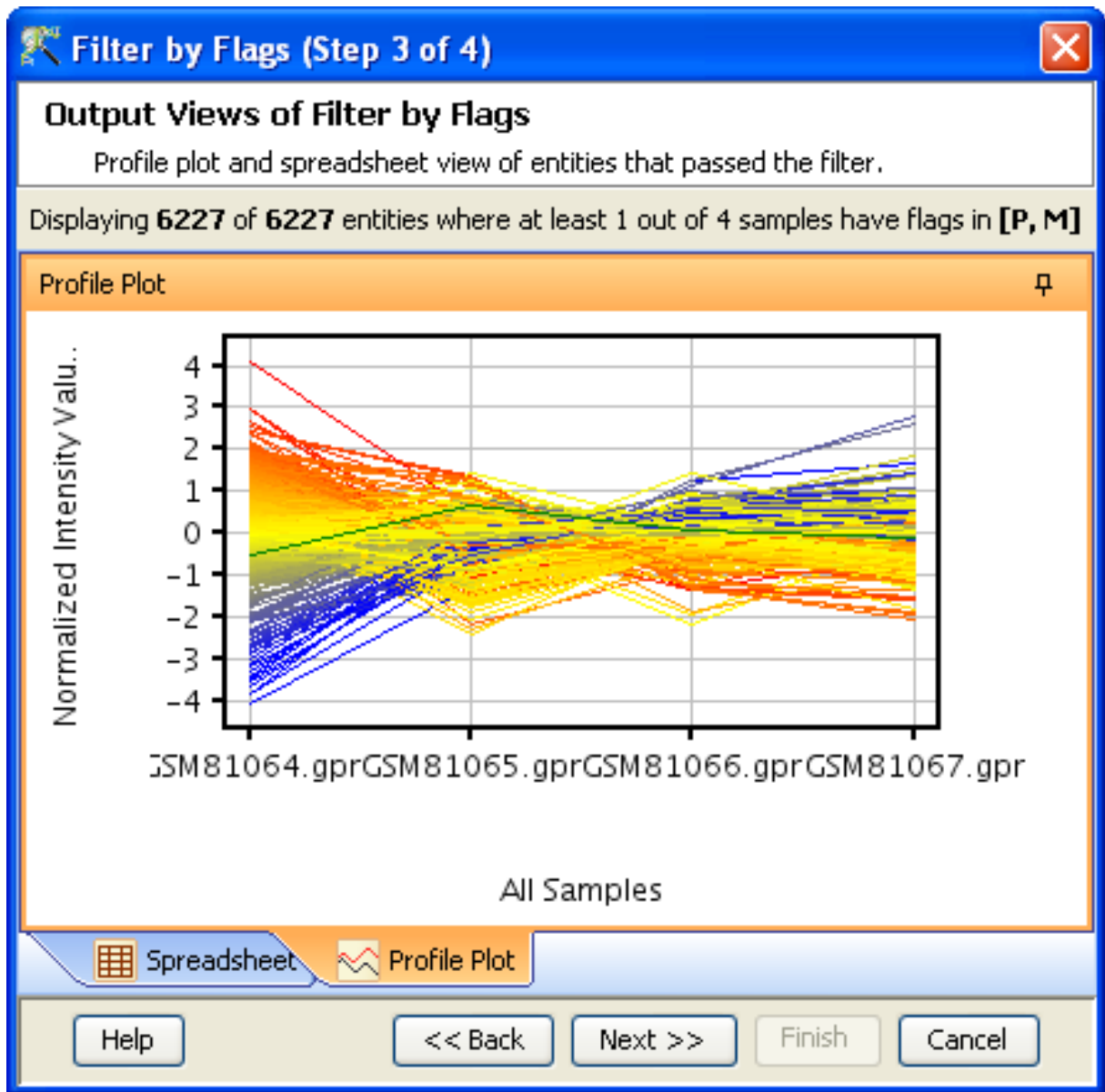


Figure 16.23: Output Views of Filter by Flags

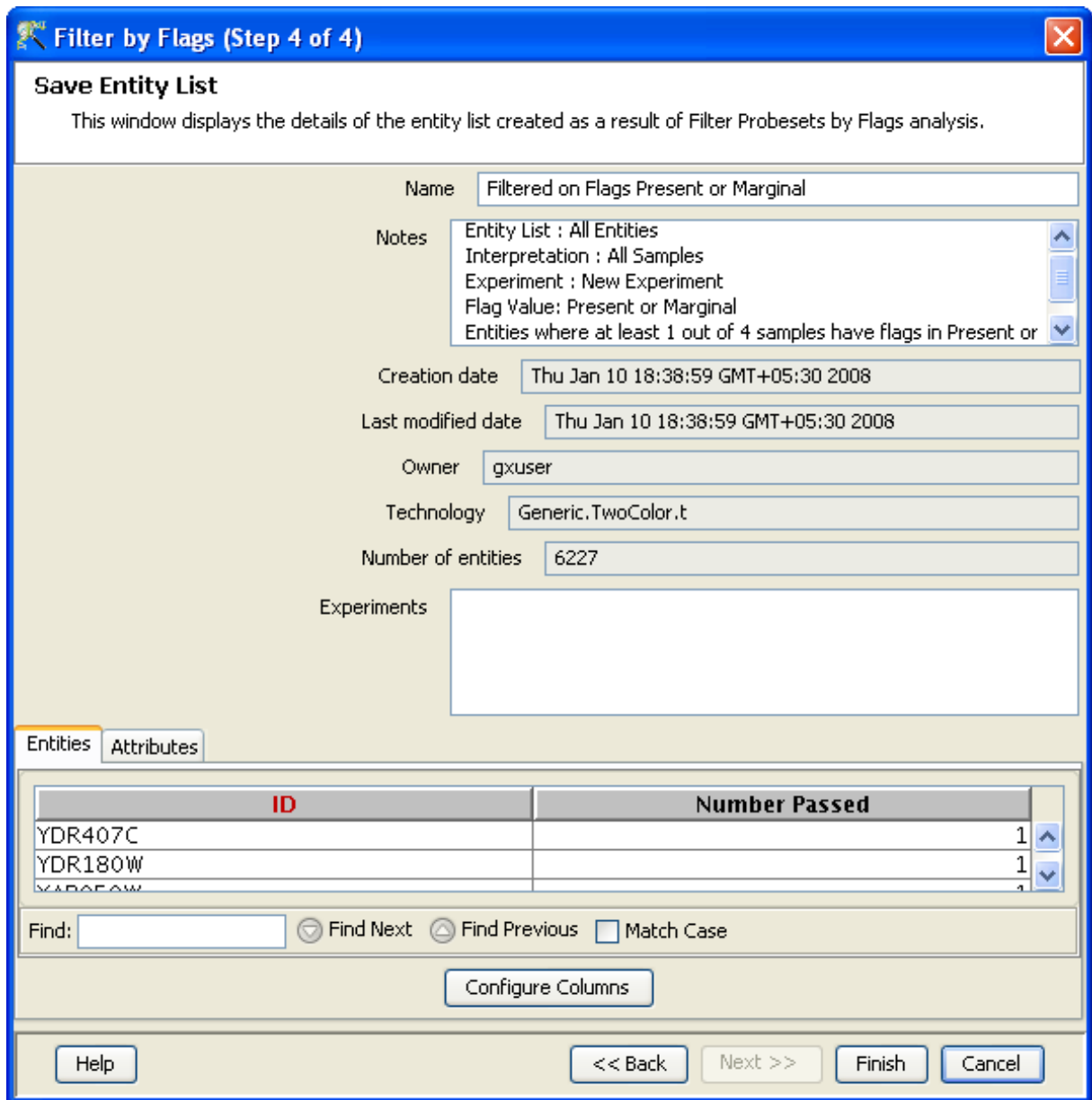


Figure 16.24: Save Entity List

- **Find Similar Entities**
For details refer to section [Find Similar Entities](#)
- **Filter on Parameters**
For details refer to section [Filter on Parameters](#)
- **Principal Component Analysis**
For details refer to section [PCA](#)

16.2.5 Class Prediction

- **Build Prediction Model** For details refer to section [Build Prediction Model](#)
- **Run Prediction** For details refer to section [Run Prediction](#)

16.2.6 Results

- **Gene Ontology (GO) analysis**
GO is discussed in a separate chapter called [Gene Ontology Analysis](#).
- **Gene Set Enrichment Analysis (GSEA)**
Gene Set Enrichment Analysis (GSEA) is discussed in a separate chapter called [GSEA](#).
- **Gene Set Analysis (GSA)**
Gene Set Analysis (GSA) is discussed in a separate chapter [GSA](#).
- **Pathway Analysis**
Pathway Analysis is discussed in a separate section called [Pathway Analysis in Microarray Experiment](#).
- **Find Similar Entity Lists**
This feature is discussed in a separate section called [Find Similar Entity Lists](#)
- **Find Significant Pathways**
This feature is discussed in a separate section called [Find Significant Pathways](#).
- **Launch IPA**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Import IPA Entity List**
This feature is discussed in detail in the chapter [Ingenuity Pathways Analysis \(IPA\) Connector](#).
- **Extract Interactions via NLP**
This feature is discussed in detail in the chapter [Pathway Analysis](#).

16.2.7 Utilities

- **Import Entity list from File** For details refer to section [Import list](#)
- **Differential Expression Guided Workflow:** For details refer to section [Differential Expression Analysis](#)
- **Filter On Entity List:** For further details refer to section [Filter On Entity List](#)
- **Remove Entities with missing signal values** For details refer to section [Remove Entities with missing values](#)

Chapter 17

Loading Experiment from NCBI GEO

17.1 Introduction

The Gene Expression Omnibus hosted at the NCBI (<http://www.ncbi.nlm.nih.gov/geo/>) is a public repository of functional genomics data sets submitted by the scientific community, with over 17,000 experiments as of Nov 2009. GeneSpring GX can import the expression data sets, directly from the main interface, by providing it with a GSE or GEO Series identifier. The data sets will be downloaded directly from the NCBI and a new experiment will be created from the data. The experimental parameters will be extracted and used to annotate the experiment.

To load the data into GeneSpring GX, first find the GSE identifier for the experiment. If the data is described in a publication, the author will probably list the GSE identifier in the material and methods or other section of the paper. If the GSE identifier is not provided, you can search for the identifier at the Gene Expression Omnibus webpage, <http://www.ncbi.nlm.nih.gov/geo/>.

17.1.1 Load a GSE dataset

To load a GEO dataset, select the menu option **Tools** → **Import NCBI GEO Experiment**. A dialog appears allowing you to enter a GSE identifier, such as gse3541. See Figure 17.1

The experiment type should be chosen from the drop down menu. Currently, only the following experiment formats are supported in their native forms:

1. Affymetrix expression
2. Agilent One Color

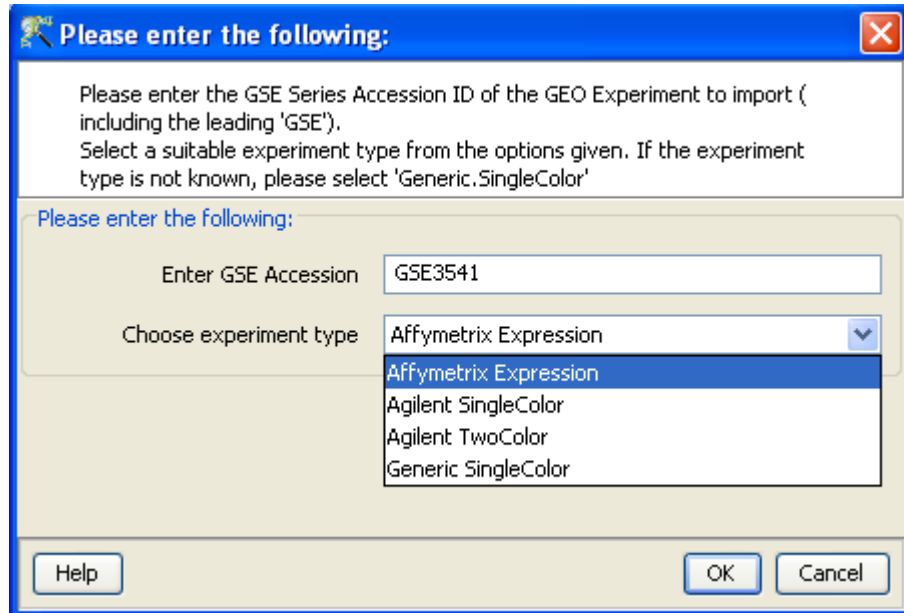


Figure 17.1: GEO Identifier Entry Dialog

3. Agilent Two Color

For Affymetrix expression data sets, only *.CEL files are supported. Pivot files are not supported in the GEO import. For Agilent expression data sets, files created with Feature Extraction version 8.5 and later are supported.

Experiments in a different format can still be loaded, but need to be loaded as 'Generic SingleColor'. This will create a technology on-the-fly (if it does not already exist) and create an experiment. If you are unsure of the experiment type, choose 'Generic SingleColor' since that will work most of the time. GeneSpring GX will also switch to importing as a 'Generic SingleColor' if the wrong experiment type is chosen (for instance, when the experiment is an Agilent experiment, but the user chooses 'Affymetrix', the experiment will be loaded as a 'Generic SingleColor' experiment.

Press *OK* to continue. The data sets will be downloaded directly from the NCBI FTP site and a progress bar will be shown.

After the data set has been successfully downloaded, a New Experiment creation window will be presented. The Experiment Name and Experiment Notes sections will be pre-populated with the information from the data set. The Experiment Name is rather long since it is based on the title of the experiment and it is suggested to choose a shorter name (although the long names will be OK in most cases).

Press *OK* to start the experiment creation. After successful creation of the experiment, the Information window will show which technology was used in the creation of the experiment and how many probes

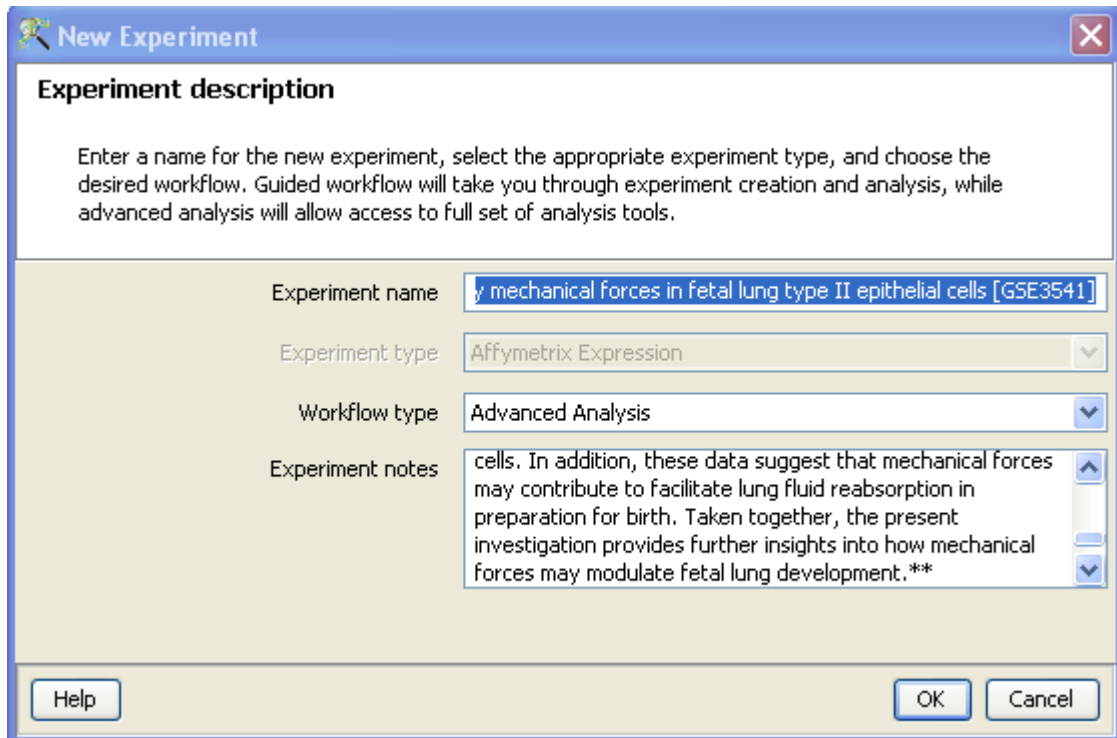


Figure 17.2: Create New Experiment Dialog

matched the probes in the technology.

17.1.2 Experiment Parameters

After an experiment is created and the data set had a corresponding GDS data set (Not guaranteed for every GSE set), the experimental grouping data is automatically copied to the experiment. Open the Experiment Grouping window from the Workflow section to see the experiment grouping information.

In this example the parameter 'stress' was copied from the GDS set 'GDS2225' and the values for each sample, such as 'control' and 'mechanical strain' are provided.

Duplicate Experiment Parameters

Sometimes, GEO creates two (or more) GDS sets from one submission GSE data set. In this case, two (or more) experiment parameters could be copied as experiment grouping parameters. An example of this is shown in Fig 17.4.

The 'agent' parameter was used in both GDS sets, but somehow the GEO curators felt that it should not

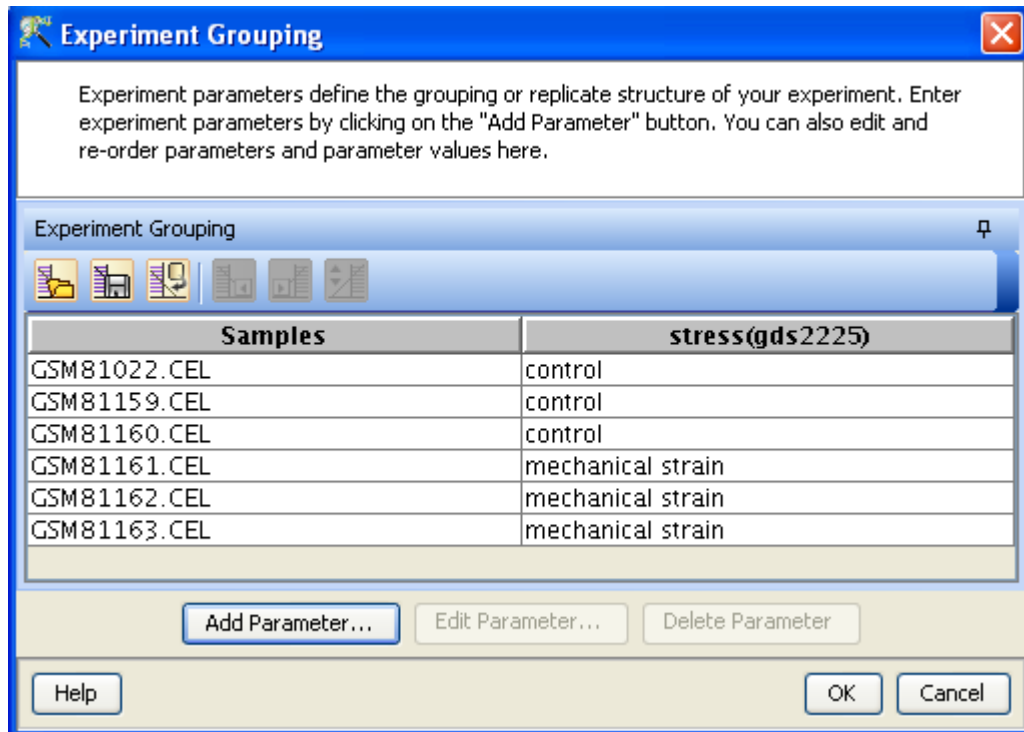


Figure 17.3: Experiment Grouping Information is automatically copied over

be the same parameter, since it was separated into two separate Geo Data Sets (GDS). In these situations it is often useful to examine the sample attributes. The sample attributes are saved with each sample and are the attributes originally submitted by the author and they may have some more information on the origin of the samples. To review the sample attributes, select the 'Import Parameters from Samples' icon in the Experiment Grouping window.

In this example it turns out that the experimental design included both shoots and roots of the Arabidopsis plant, something that was not recorded in the GDS sets.

The sample attribute 'Source' (Source_name_ch1) seems a good candidate for an experimental parameter and should be added to the experiment grouping for this experiment, by selecting the column and pressing 'Add'. Some editing of the values is required, to make these useful parameters. The two 'agent' parameters can then be combined into one parameter 'agent' and the experiment is now ready for further analysis.

17.2 Possible Error Messages

Invalid GSEid provided: GeneSpring GX currently only accepts the GSE identifiers. The GSE identifiers represent the original dataset submitted by the author. If an invalid identifier is used, the error message *Do not enter GSD or other GEO accession numbers* is displayed.

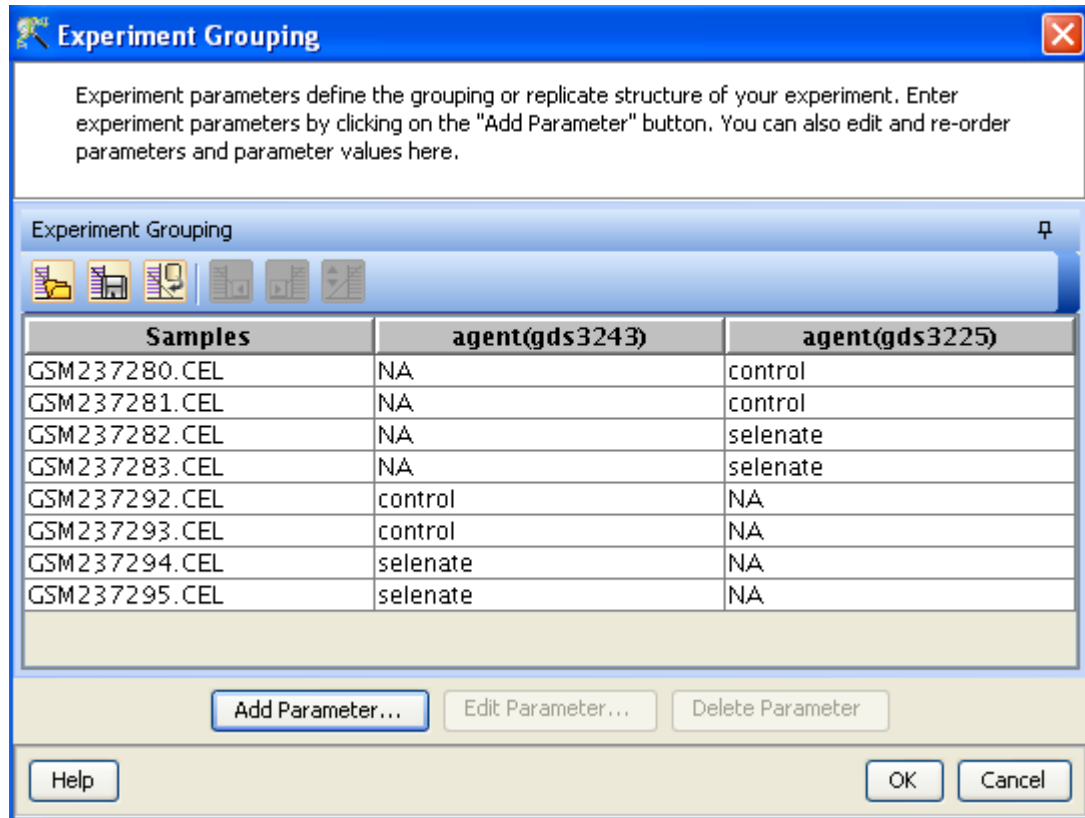


Figure 17.4: Duplicate Experiment Parameters

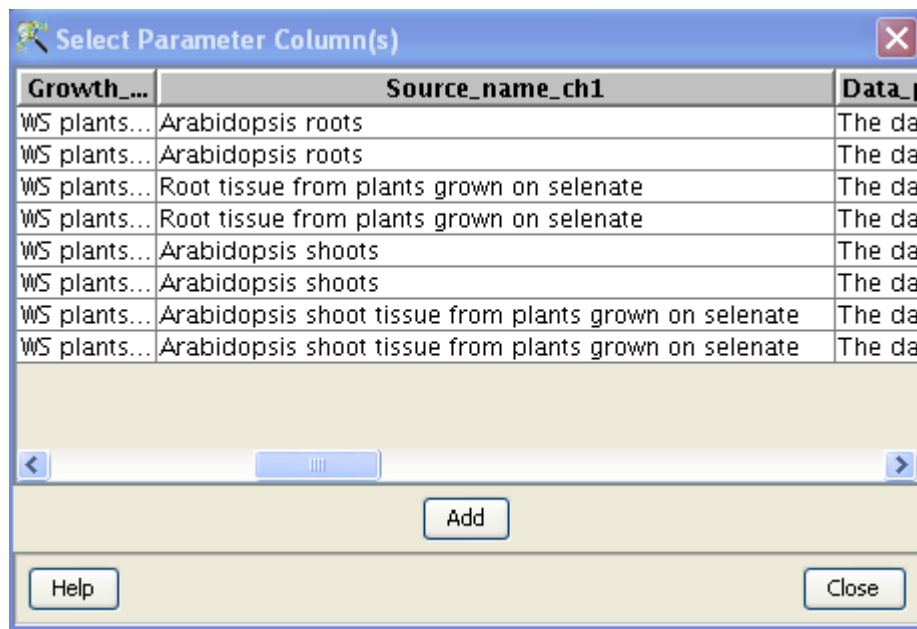


Figure 17.5: Duplicate Parameters

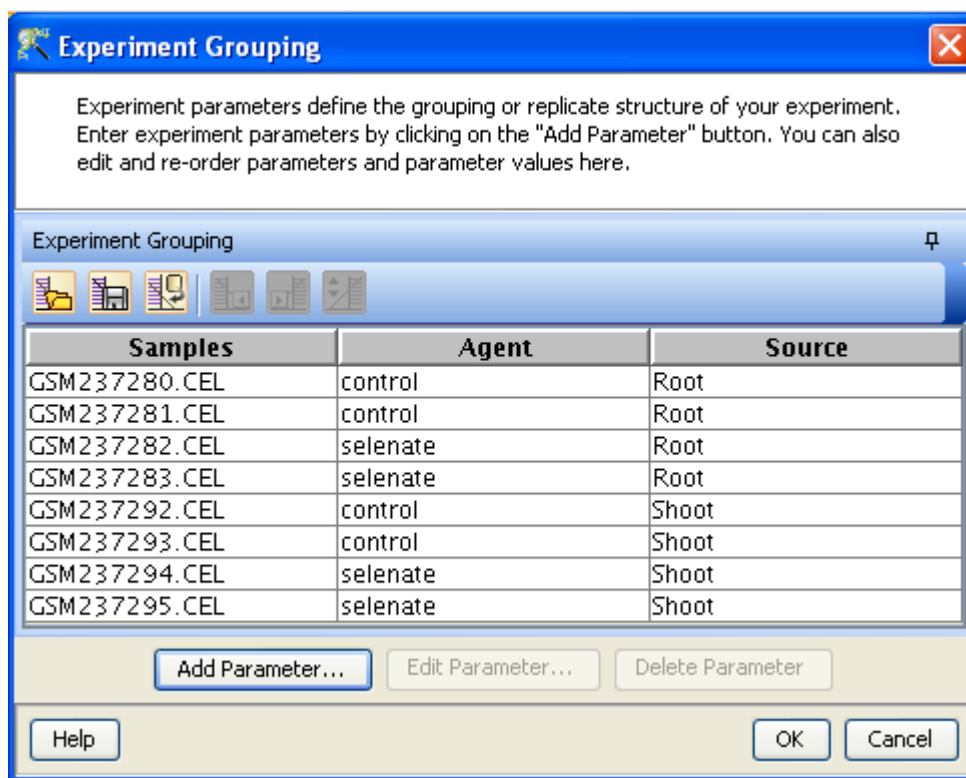


Figure 17.6: Final Experiment Grouping

Technology type does not match: When an incorrect experiment type was selected in the first dialog box, a warning dialog may be presented. This will happen most likely if the GSE data set is NOT an Affymetrix data set, but something else. GeneSpring GX will still be able to import the data set as a Generic Single Color experiment. Press *YES* to import the experiment as a generic single color experiment or press *NO* to cancel the import.

If the GSE data set is an Affymetrix data set and the user chooses 'Agilent Single (or Two) color' in the experiment type, this window will not appear, but the Sample Chooser 'New Experiment (Step 1 of 5)' will be empty. This is an indication that the experiment type is not an Agilent experiment. Cancel the experiment creation and execute the GEO importer again, choosing 'Affymetrix' or 'Generic SingleColor' as the experiment type. The experiment type should also be obvious from the information on the GSE data set on the GEO website. Consult the GEO website for more information on the chosen GSE data set if unsure about the origin of the data set.

Supplementary files are unavailable: Some GSE data sets do not have any of the original data files for the experiment. When the user chooses 'Affymetrix' or 'Agilent' as the experiment type and the GSE data set does not contain any original data files, a warning dialog is shown.

GeneSpring GX will still be able to import the data set as a Generic Single Color experiment. Press *YES* to import the experiment as a generic single color experiment or press *NO* to cancel the import.

Experiment creation failed: This message may appear when no connection can be made to the GEO FTP site or other network problems. Check the network connection. GeneSpring GX attempts to

connect to the FTP site at the <ftp://ftp.ncbi.nlm.gov>. Check to see if the FTP site can be reached with other tools.

Unable to validate Experiment information: If this window appears, the most likely reason is the fact that the GSE data set contains data for two different chips or technologies. At this point only data sets from a single technology (chip) can be loaded directly from GEO.

17.3 Experiment Parameters and Sample Attributes

Experiment parameters, such as treatment, source etc. are usually not part of the GSE data sets as submitted by the research community. The experiment parameters are created by the curation staff of GEO and are part of the GEO Data Set or GDS. These GDS are not available for every data set that is submitted to GEO. The curation is a complicated and time-consuming process and the GEO staff is currently experiencing a considerable backlog. Also, not every data set submission to GEO is suitable for curation. Therefore, it is not guaranteed that GeneSpring GX will be able to extract the Experiment Parameters for every GSE data set that is available on GEO.

17.3.1 Create Experiment Parameters from Sample Attributes

During the download of the GSE data sets, each sample from the data set is annotated with a number of attributes. These attributes are usually provided directly by the submitter and sometimes (but not always!) contains information on the experimental design or experimental parameters.

These sample attributes can potentially be used as experiment parameters. The Experiment Grouping window will allow one or more columns to be used as samples attributes. Open the Experiment Grouping window from the Workflow section **Experiment Setup** and click on **Import Parameters from Samples** icon.

This window will show all the sample attributes. Many of the sample attributes are the same for all samples, such as 'Platform' or 'Submission Date' and would not be useful experiment parameters, but 'Title' usually contains some indication of the experimental conditions that are important for the analysis (such as 'Myoblast (1) G1' and 'Myotube (D1)' in the example below).

The actual value of the 'Title' column would not be suitable as an Experiment Grouping parameter, since none of the values for an experimental condition are the same, but the column can be loaded as a parameter and later edited in the Experimental Grouping window.

Choosing the 'Title' as the Experiment Parameter for this experiment makes it easier to edit the values to their proper value and avoids errors. Select *Add* to add the selected column 'Title' as an experiment parameter and edit the parameter values to contain correct experiment parameter values that can be used in the creation of Interpretations and perform statistical analysis.

The image shows a dialog box titled "Select Parameter Column(s)" with a close button in the top right corner. The dialog contains a table with the following data:

Geo_acce...	Contact_...	Title	Label_ch1	Last
GSM119...	University...	NormalControl, rep1	Biotin	Apr 1
GSM119...	University...	NormalControl, rep2	Biotin	Apr 1
GSM119...	University...	RA, rep1, pre-treatment	Biotin	Apr 1
GSM119...	University...	RA, rep2, pre-treatment.	Biotin	Apr 1
GSM119...	University...	RA, rep1, post-treatment	Biotin	Apr 1
GSM119...	University...	RA, rep2, post-treatment.	Biotin	Apr 1

Below the table is a horizontal scrollbar. At the bottom of the dialog, there are three buttons: "Add", "Help", and "Close".

Figure 17.7: Sample attributes that can be chosen as Experiment Parameters

Chapter 18

Advanced Workflow

The *Advanced Workflow* in **GeneSpring GX** provides tremendous flexibility and power to analyze your microarray data depending upon the technology used, the experimental design and the focus of the study. *Advanced Workflow* provides several choices in terms of summarization algorithms, normalization routines, baseline transform options and options for flagging spots depending upon the technology. All these choices are available to the user at the time of experiment creation. The choices are specific for each technology (Agilent, Affymetrix, Illumina and Generic Technologies) and are described under the *Advanced Workflow* section of the respective chapters. Additionally, *Advanced Workflow* also enables the user to create different interpretations to carry out the analysis. Other features exclusive to *Advanced Workflow* are options to choose the p-value computation methods (Asymptotic or permutative), p-value correction types (e.g., Benjamini-Hochberg or Bonferroni), Principal component Analysis (PCA) on the entities, Class Prediction, Gene Set Enrichment Analysis (GSEA), Importing BioPax pathways and several other utilities. The *Advanced Workflow* can be accessed by choosing *Advanced* as the *Workflow Type*, in the *New Experiment* box, at the start of the experiment creation. If the experiment has been created in a *Guided* mode, then the user does not have the option to choose the summarization, normalization and baseline transformation, i.e. the experiment creation options. However, one can still access the analysis options available from the *Advanced Workflow*, which opens up after the experiment is created and preliminary analysis done in Guided mode.

Described below are the sections of the *Advanced Workflow*:

18.1 Experiment Setup

18.1.1 Quick Start Guide

Clicking on this link will take you to the appropriate chapter in the on-line manual giving details about: loading expression files into **GeneSpring GX** , *Advanced Workflow*, the method of analysis, the details

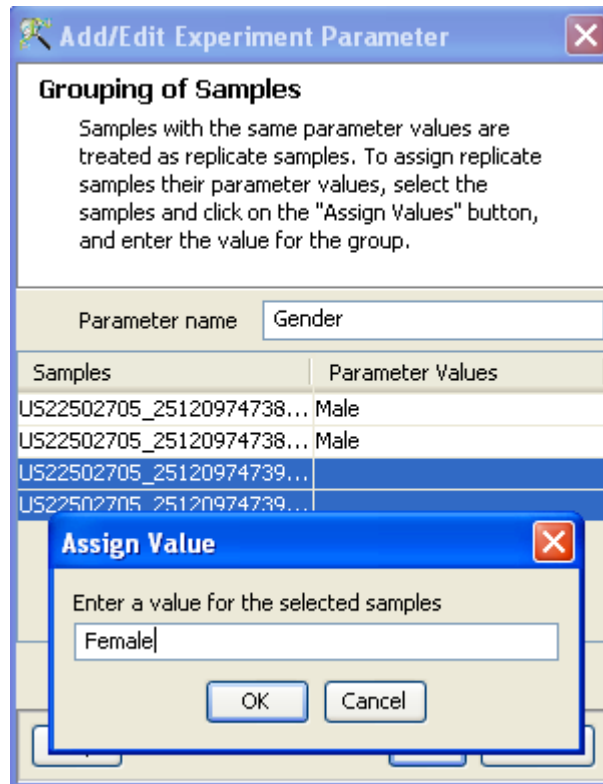




Figure 18.1: Experiment Grouping

of the algorithms used and the interpretation of results.

18.1.2 Experiment Grouping





Experiment Grouping requires the adding of parameters to help define the grouping and replicate structure of the experiment. Parameters can be created by clicking on the *Add parameter* button. Sample values can be assigned by first selecting the desired samples and assigning the value. For removing a particular value, select the sample and click on *Clear*. Press *OK* to proceed. Any number of parameters can be added for analysis in the *Advanced Analysis*.

Experimental parameters can also be loaded, using Load experiment parameters from file  icon, from a tab or comma separated text file, containing the *Experiment Grouping* information. The experimental parameters can also be imported from previously used samples, by clicking on Import parameters from samples  icon. In case of file import, the file should contain a column containing sample names; in addition, it should have one column per factor containing the grouping information for that factor. Here is an example of a tab separated file.

Sample genotype dosage

A1.txt NT 20
A2.txt T 0
A3.txt NT 20
A4.txt T 20
A5.txt NT 50
A6.txt T 50

Reading this tab file generates new columns corresponding to each factor.

The current set of newly entered experiment parameters can also be saved in a tab separated text file, using Save experiment parameters to file  icon. These saved parameters can then be imported and re-used for another experiment as described earlier. In case of multiple parameters, the individual parameters can be re-arranged and moved left or right. This can be done by first selecting a column by clicking on it and using the Move parameter left  icon to move it left and Move parameter right  icon to move it right. This can also be accomplished using the Right click \rightarrow *Properties* \rightarrow columns option. Similarly, parameter values, in a selected parameter column, can be sorted and re-ordered, by clicking on Re-order parameter values  icon. Sorting of parameter values can also be done by clicking on the specific column header.

Unwanted parameter columns can be removed by using the Right-click \rightarrow *Properties* option. The *Delete parameter* button allows the deletion of the selected column. Multiple parameters can be deleted at the same time. Similarly, by clicking on the *Edit parameter* button the parameter name as well as the values assigned to it can be edited.

18.1.3 Create Interpretation

An interpretation specifies how the samples should be grouped into experimental conditions. the interpretation can be used for both visualization and analysis. Interpretation can be created using the *Create interpretation* wizard which involves the following steps:

Step 1 of 3: Experiment parameters are shown in this step. In case of multiple parameters, all the parameters will be displayed. The user is required to select the parameter(s) using which the interpretation is to be created.

Step 2 of 3: Allows the user to select the conditions of the parameters which are to be included in the interpretation. All the conditions (including combinations across the different parameters) are shown. By default all these experimental conditions are selected, click on the box to unselect any. Any combination of these conditions can be chosen to form an interpretation. If there are multiple samples for a condition, users can use average over these samples by selecting the option *Average over replicates in conditions* provided at the bottom of the panel. Please note that all analysis do not use the average of replicates. For example, while performing statistical analysis the interpretation that is used is always the non averaged interpretation. So even if the interpretation selected is averaged, the tool considers it as unaveraged.

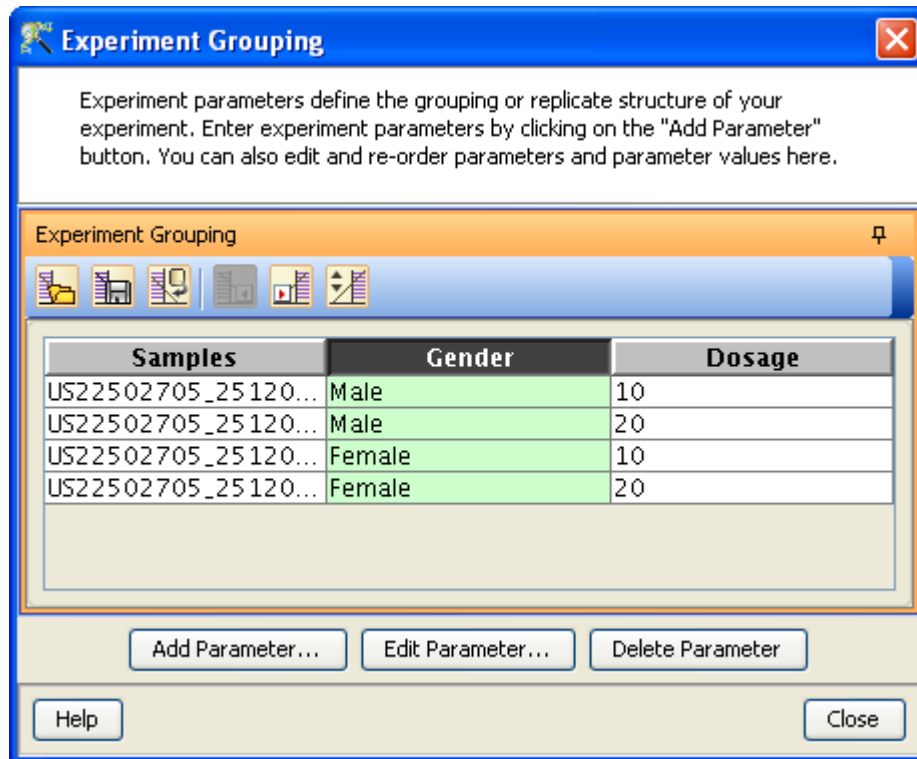


Figure 18.2: Edit or Delete of Parameters

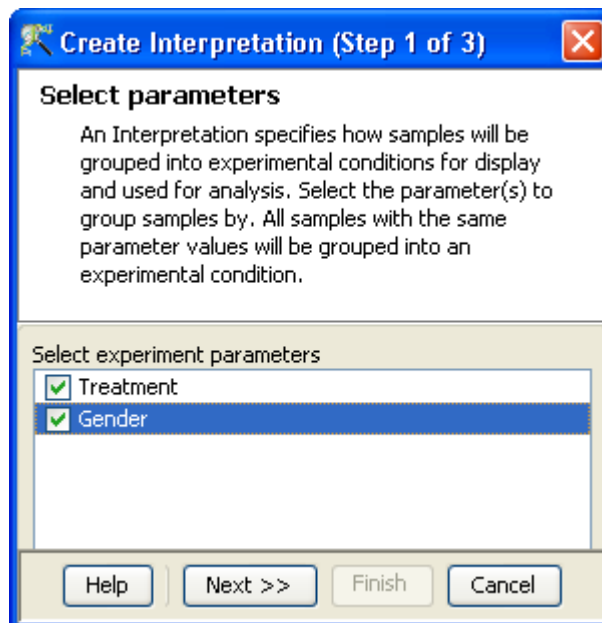


Figure 18.3: Create Interpretation (Step 1 of 3)

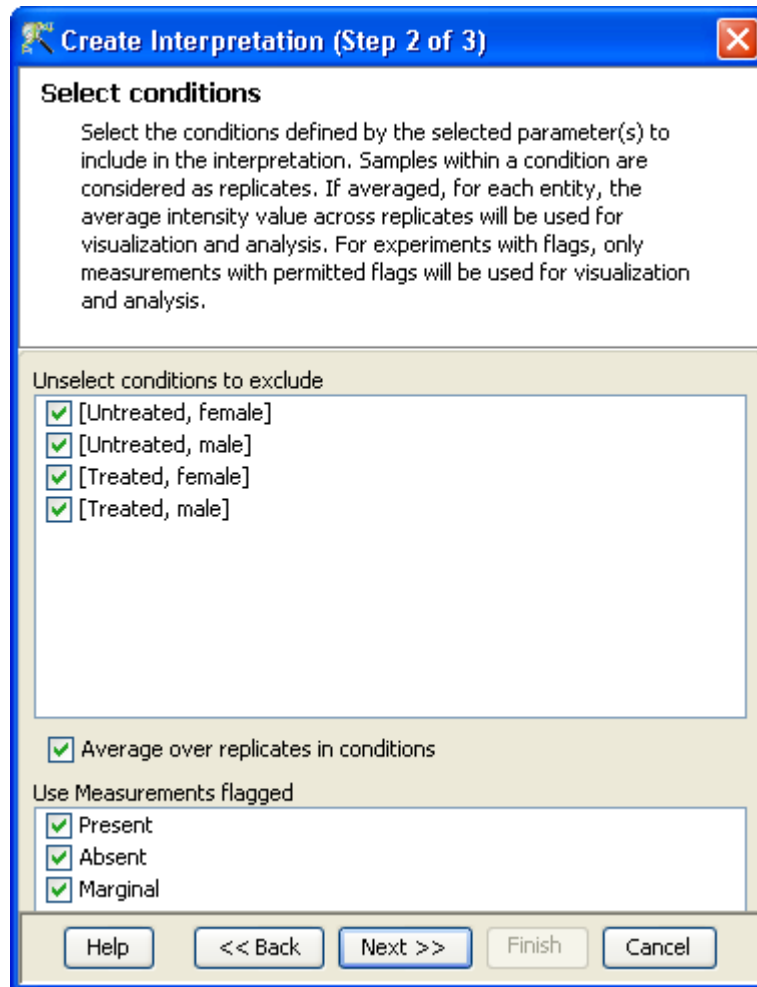


Figure 18.4: Create Interpretation (Step 2 of 3)

The user also has an option to exclude entities with flags while creating an interpretation. By default all the flags are included in the interpretation and in case the user wants to exclude any flags, he/she can unselect the same during the interpretation. The user can see the effect of this interpretation on the entity list by viewing the spreadsheet. The effect of excluding flag values on all the samples can be seen by viewing the unaveraged interpretation. This shows the entities and the values which have been excluded, appear blank. The spreadsheet can be viewed by selecting the desired entity list and the interpretation (the selected interpretation appears in bold and the selected entity list is highlighted). However analysis such as clustering, class prediction and PCA take all the flags into account even if specified otherwise in the interpretation.

For more information, on the effect of interpretation on the analysis as well as the way the interpretations are handled in different analysis refer to the section on [Conditions and Interpretations](#).

Step 3 of 3: This page displays the details of the interpretation created. This includes user editable Name for the interpretation and Notes for description of the interpretation. Descriptions like creation date, last modification date, and owner are also present, but are not editable.

Figure 18.5: Create Interpretation (Step 2 of 3)

18.1.4 Create new Gene Level Experiment

Create new gene level experiment is a utility in **GeneSpring GX** that allows analysis at gene level, even though the signal values are present only at probe level. Suppose an array has 10 different probe sets corresponding to the same gene, this utility allows summarizing across the 10 probes to come up with one signal at the gene level and use this value to perform analysis at the gene level.

Process

- *Create new gene level experiment* is supported for all those technologies where gene Entrez ID column is available. It creates a new experiment with all the data from the original experiment; even those probes which are not associated with any gene Entrez ID are retained.
- The identifier in the new gene level experiment will be the Probe IDs concatenated with the gene entrez ID; the identifier is only the Probe ID(s) if there was no associated entrez ID.
- Each new gene level experiment creation will result in the creation of a new technology on the fly.
- The annotation columns in the original experiment will be carried over except for the following.

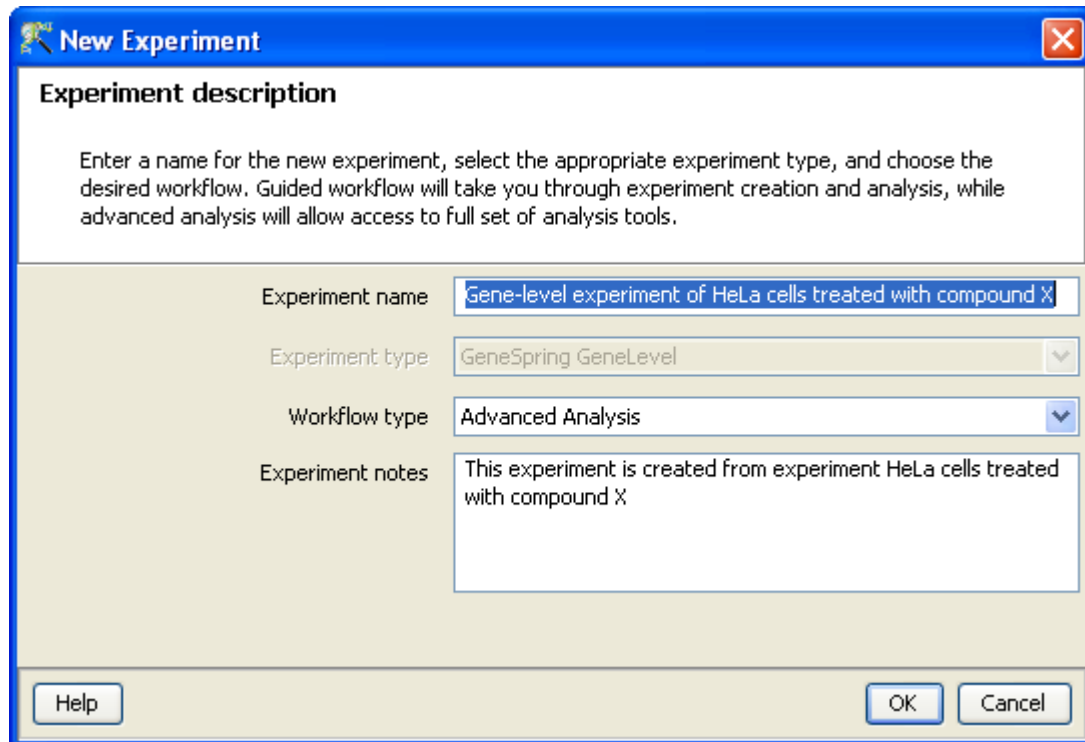


Figure 18.6: Gene Level Experiment Creation

- Chromosome Start Index
 - Chromosome End Index
 - Chromosome Map
 - Cytoband
 - Probe Sequence
- Flag information will also be dropped.
 - Raw signal values are used for creating gene level experiment; if the original experiment has raw signal values in log scale, the log scale is retained.
 - Experiment grouping, if present in the original experiment, will be retained.
 - The signal values will be averaged over the probes (for that gene entrez ID) for the new experiment.

Create new gene level experiment can be launched from the **Workflow Browser** → **Experiment Set up**. An experiment creation window opens up; experiment name and notes can be defined here. Note that only advanced analysis is supported for gene level experiment. Click *OK* to proceed.

A three-step wizard will open up.

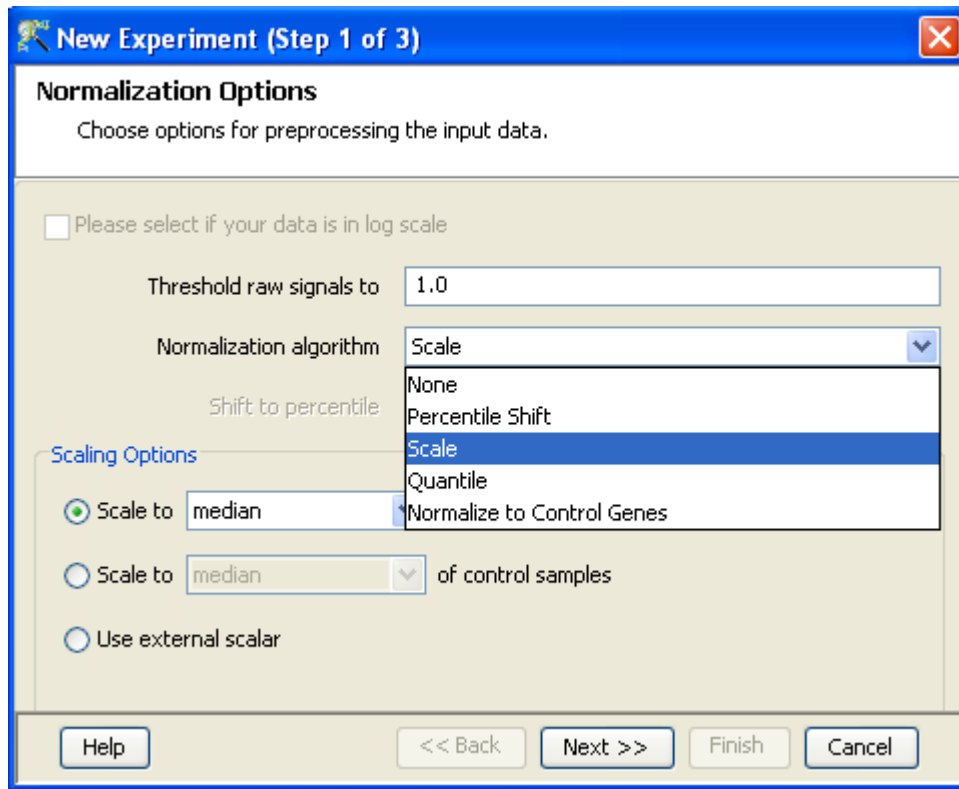


Figure 18.7: Gene Level Experiment Creation - Normalization Options

Step 1: Normalization Options If the data is in log scale, the thresholding option will be greyed out.

Normalization options are:

- **None:** Does not carry out normalization.
- **Percentile Shift:** On selecting this normalization method, the **Shift to Percentile Value** box gets enabled allowing the user to enter a specific percentile value.
- **Scale:** On selecting this normalization method, the user is presented with an option to either scale it to the median/mean of all samples or to scale it to the median/mean of control samples. On choosing the latter, the user has to select the control samples from the available samples in the **Choose Samples** box. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.
- **Quantile:** Will make the distribution of expression values of all samples in an experiment the same.
- **Normalize to control genes:** After selecting this option, the user has to specify the control genes in the next wizard. The **Shift to percentile** box is disabled and the percentile is set at a default value of 50.

See Chapter [Normalization Algorithms](#) for details on normalization algorithms.

Step 2: Choose Entities If the **Normalize to control genes** option is chosen in the previous step, then the list of control entities can be specified in the following ways in this wizard:

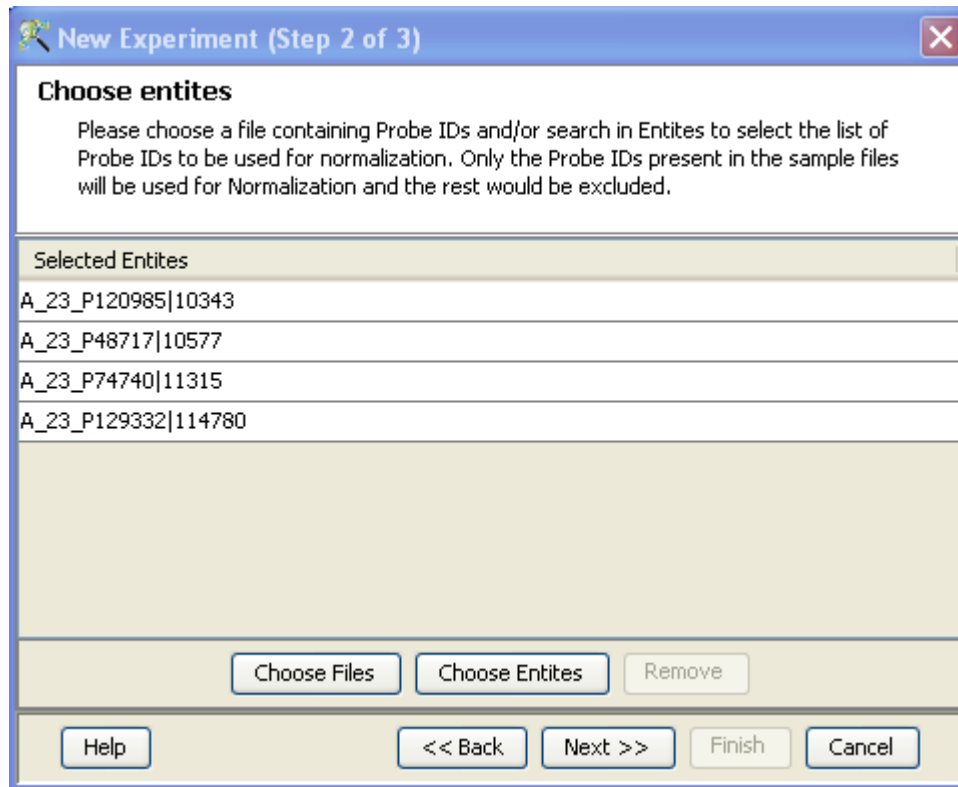


Figure 18.8: Gene Level Experiment Creation - Choose Entities

- By choosing a file(s) (txt, csv or tsv) which contains the control entities of choice denoted by their probe id. Any other annotation will not be suitable.
- By searching for a particular entity by using the *Choose Entities* option. This leads to a search wizard in which the entities can be selected. All the annotation columns present in the technology are provided and the user can search using terms from any of the columns. The user has to select the entities that he/she wants to use as controls, when they appear in the **Output Views** page and then click *Finish*. This will result in the entities getting selected as control entities and will appear in the wizard.

The user can choose either one or both the options to select his/her control genes. The chosen genes can also be removed after selecting the same.

In case the entities chosen are not present in the technology or sample, they will not be taken into account during experiment creation. The entities which are present in the process of experiment creation will appear under matched probe IDs whereas the entities not present will appear under unmatched probe ids in the experiment notes in the experiment inspector.

Step 3: Preprocess Baseline Options This step allows defining base line transformation operations. Click *Ok* to finish the gene level experiment creation.

A new experiment titled "Gene-level experiment of original experiment" is created and all regular analysis possible on the original experiment can be carried out here also.

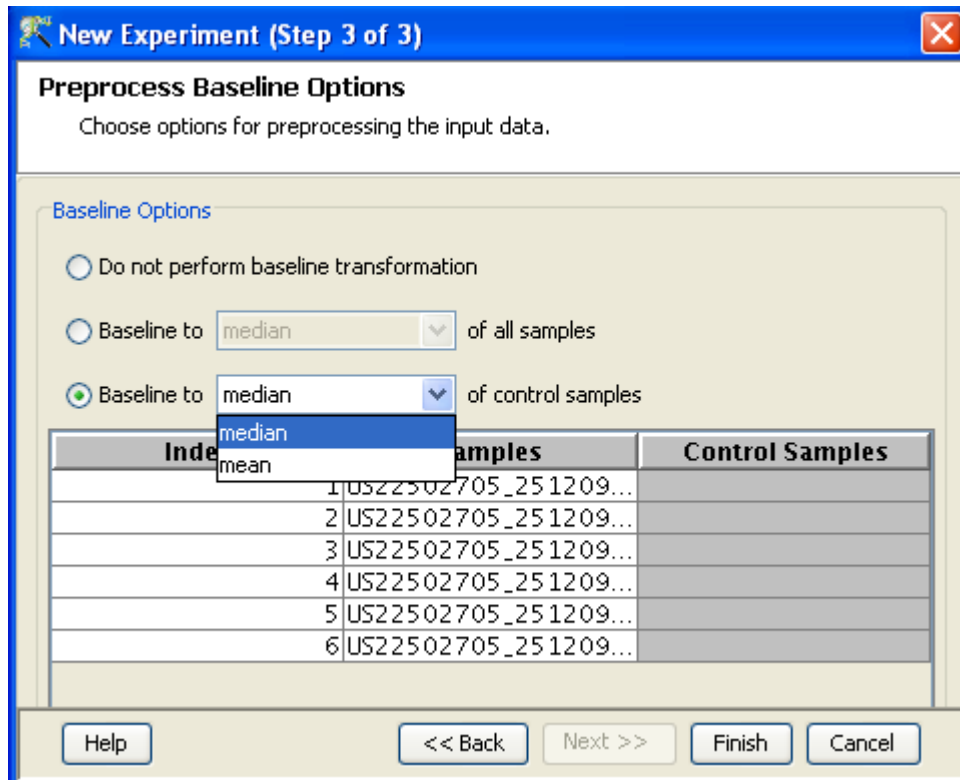


Figure 18.9: Gene Level Experiment Creation - Preprocess Baseline Options

18.2 Quality Control

18.2.1 Quality Control on Samples

Quality control is an important step in micro array data analysis. The data needs to be examined and ambiguous samples should be removed before starting any data analysis. Since microarray technology is varied, quality measures have to be vendor and technology specific. **GeneSpring GX** packages vendor and technology specific quality measures for quality assessment. It also provides rich, interactive and dynamic set of visualizations for the user to examine the quality of data. Details of the QC metric used for each technology can be accessed by clicking on the links below.

- [Quality Control for Affymetrix Expression](#)
- [Quality Control for Exon Expression](#)
- [Quality Control for Exon Splicing](#)
- [Quality Control for Agilent Single Color](#)
- [Quality Control for Agilent Two Color](#)

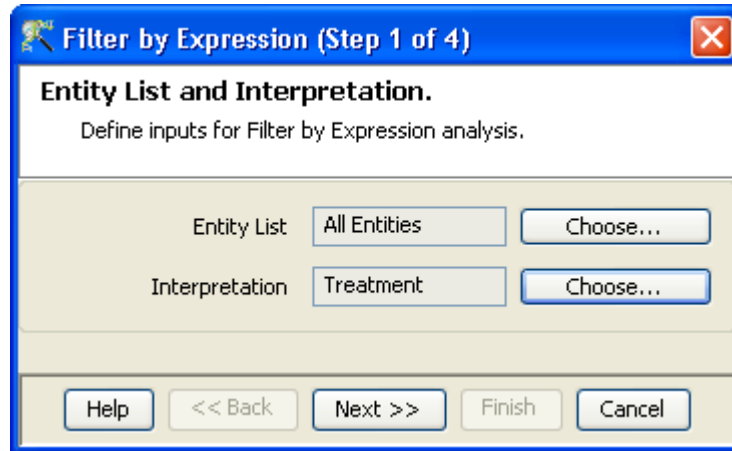


Figure 18.10: Filter probesets by expression (Step 1 of 4)

- [Quality Control for Agilent miRNA](#)
- [Quality Control for Illumina](#)
- [Quality Control for Generic Single Color](#)
- [Quality Control for Generic Two Color](#)
- [Quality Control for RealTime PCR](#)

18.2.2 Filter Probesets by Expression

Entities are filtered based on their signal intensity values. This enables the user to remove very low signal values or those that have reached saturation. Users can decide the proportion of conditions must meet a certain threshold. The *Filter by Expression* wizard involves the following 4 steps:

Step 1 of 4: Entity list and the interpretation on which filtering is to be done is chosen in this step. Click *Next*.

Step 2 of 4: This step allows the user to select the range of intensity value within which the probe intensities should lie. By lowering the upper percentile cutoff from 100%, saturated probes can be avoided. Similarly increasing the lower percentile cut off, probes biased heavily by background can be excluded. Stringency of the filter can be set in *Retain Entities* box. These fields allow entities that pass the filtering settings in some but not all conditions to be included in the filter results.

With two dye experiments, there are actually 2 values per entity per sample. When *Filter by Expression* is carried out on raw data with two-dye experiments, note that an entity is included in filtered results if *either or both* of the channels pass the defined cut-off.

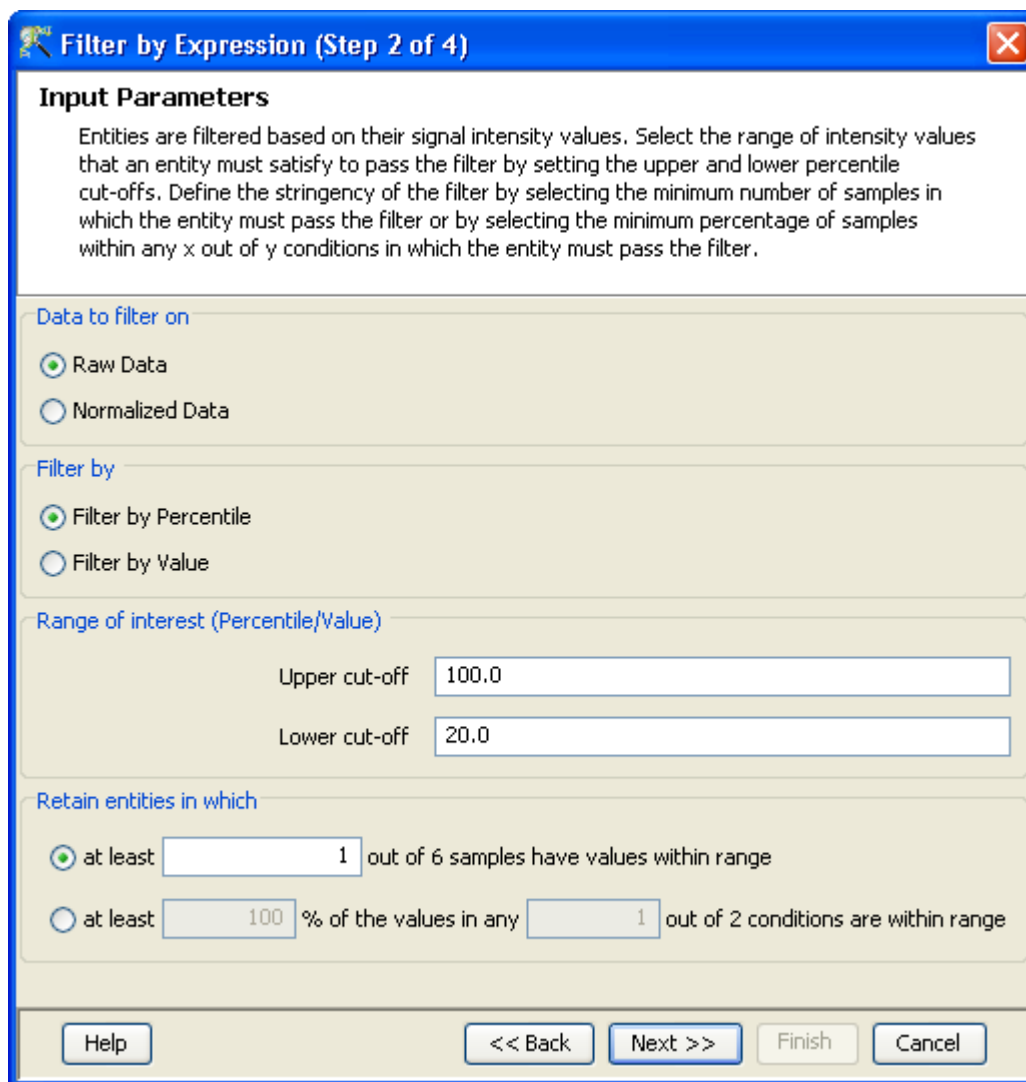


Figure 18.11: Filter probesets by expression (Step 2 of 4)

Step 3 of 4: This window shows the entities which have passed the filter, in the form of a spreadsheet and a profile plot. Number of entities passing the filter is mentioned at the top of the panel. Click *Next*.

Step 4 of 4 The last page shows all the entities passing the filter along with their annotations. It also shows the details (regarding Creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to entities which satisfied the cutoff. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using *Configure Columns* button.

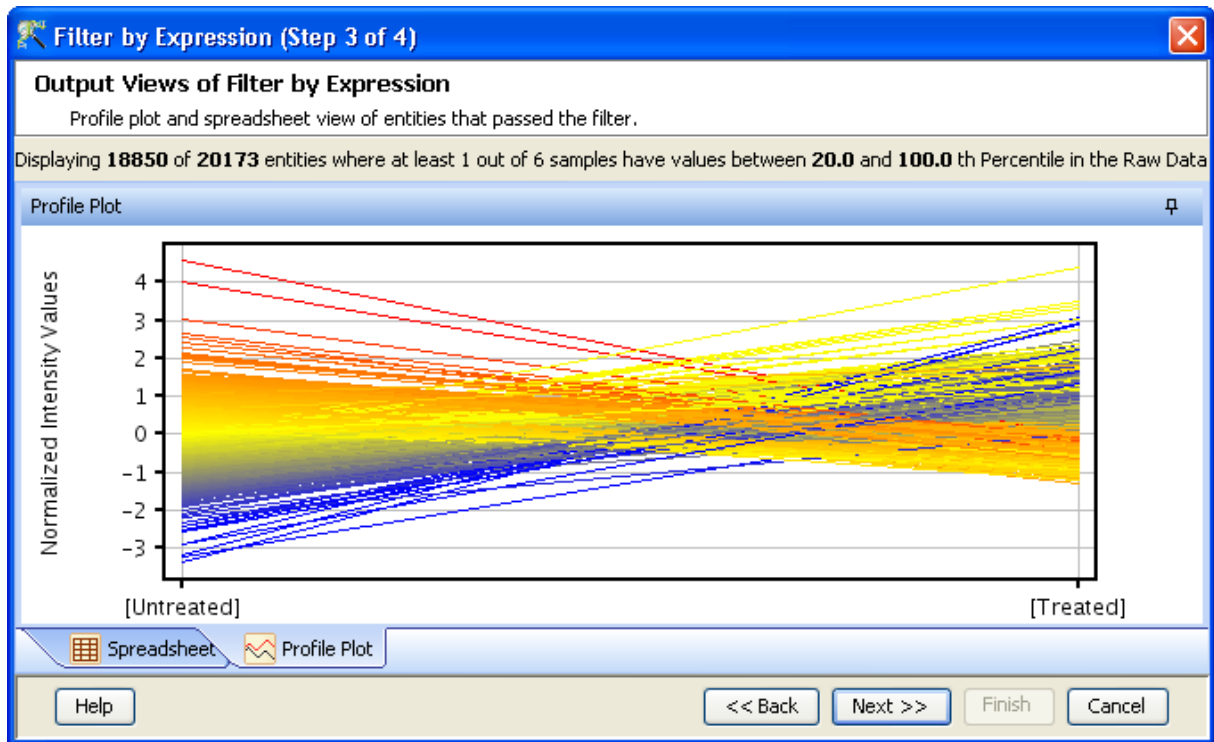


Figure 18.12: Filter probesets by expression (Step 3 of 4)

18.2.3 Filter probesets by Flags

Flags are attributes that denote the quality of the entities. These flags are generally specific to the technology or the array type used. Thus the experiment technology type, i.e., Agilent Single Color, Agilent Two Color, Affymetrix Expression, Affymetrix Exon Expression, and Illumina Bead technology determine the flag notation. These technology specific flags are described in the respective technology specific section.

For details refer to sections

- [Filter probesets for Affymetrix expression](#)
- [Filter probesets for Exon expression](#)
- [Filter probesets for agilent single color](#)
- [Filter probesets for agilent two color](#)
- [Filter probesets for illumina](#)
- [Filter probesets for generic single color](#)
- [Filter probesets for generic two color](#)

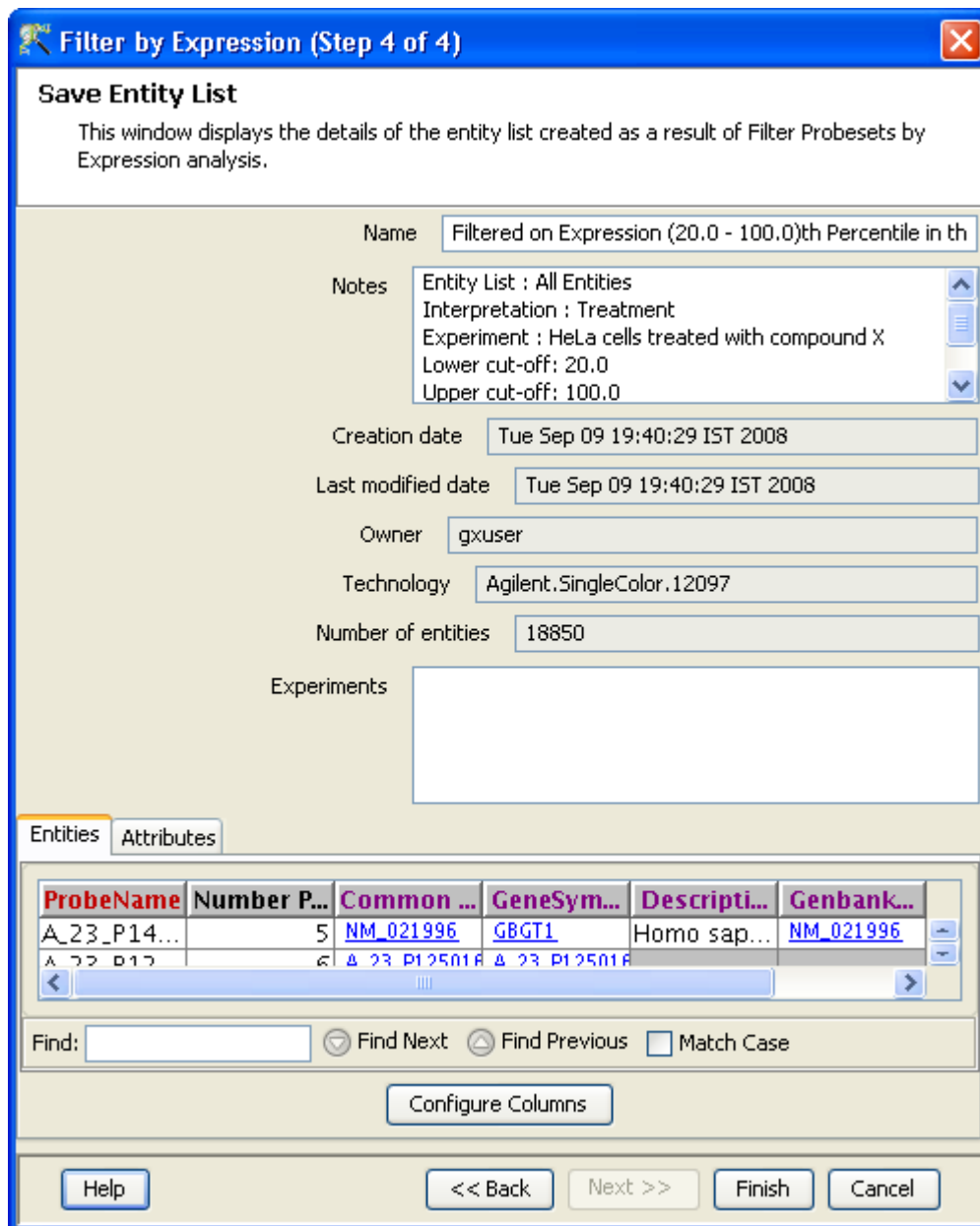


Figure 18.13: Filter probesets by expression (Step 4 of 4)

18.2.4 Filter Probesets on Data Files

The *Filter Probesets on Data Files* link is present under the *Quality Control* options in the Advanced workflow. This link allows the user to filter entities based on values in a specific column of your original data files. This filter lets you choose any of the columns in your data file and filter on the contents, both numeric and character data.

After selecting this option, the user has to go through the following steps:

- The **Input Parameters** window allows the selection of an entity list and an interpretation on which the filtering needs to be performed. This is enabled by selecting the Choose button which shows all the available entity lists and interpretations.
- The **Sample Preview** window shows the first 100 rows of the first sample (as all samples belonging to the same experiment have the same file format) and gives an idea about the columns present in the sample file and their content type. The condition panel allows adding one or more conditions for searching on samples and also to combine search conditions using either AND/OR. The search field shows a dropdown with all the column names and depending on the content of the column (numeric or character) the options for the condition changes. The Search value that needs to be filled up should be the one that is common to the entities of interest. The stringency of the filter can be set in *Retain Entities* box.
- In the **Output Views** window, a spreadsheet and a profile plot appear as two tabs, displaying those probes which have passed the filter conditions. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. The visualization shows the values after data processing (normalization, baseline transformation etc).
- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding Creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using *Configure Columns* button. Selecting Finish results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator.

18.2.5 Filter Probesets by Error

This option allows the user to filter on the standard deviation or the coefficient of variation (CV). The option to filter on standard deviation or CV allows the user to filter entities which are above or below the value specified by the user. The user can filter on standard deviation among groups in case the standard deviation is comparable between the groups or he/she can filter on % CV if the standard deviation between the groups is highly varied.

In other words, filtering by CV renders the comparison of standard deviation, mean insensitive. If the condition specified is greater than equal to, then all the entities having values greater than or equal to the specified value are retained and in case the condition specified is lesser than, then the entities having

values lesser than the specified value are retained for further analysis. This filtering option can be used for achieving two kinds of objectives:

1. To filter out genes having outlier samples
2. To filter out genes having low variation in expression values across all samples or in an extreme case-constant values (This can be done by choosing the interpretation All Samples)

After selecting the *Filter on Error* option, the user has to go through the following steps:

- The **Entity list** and **Interpretation** window allows the selection of an entity list and an interpretation. This is enabled by selecting the *Choose* button which shows all the available entity lists and interpretations. The unaveraged interpretation is always considered for this analysis.
- The **Input Parameters** window allows the selection of either standard deviation or CV as the filtering option. It also allows the stringency of the filter to be set in the *Retain Entities* box.
- In the **Output Views** window, a spreadsheet and a profile plot appear as two tabs, displaying those probes which have passed the filter conditions. Total number of probes and number of probes passing the filter are displayed on the top of the navigator window. The profile plot shows the processed data values and the spreadsheet shows all the entities along with the number of conditions in which they passed the filter criteria and either the CV or the standard deviation values.
- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding Creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using *Configure Columns* button. Selecting *Finish* results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator.

18.3 Analysis

18.3.1 Statistical Analysis

A variety of statistical tests are available depending on the experimental design. The *Statistical Analysis* wizard has 9 steps, which are selectively shown based on the input. Using the experimental design given in table 18.1 as an example, the steps involved in the wizard are described below. This particular experimental design would use t-test for the analysis.

Step 1 of 9: Entity list and the interpretation on which analysis is to be done is chosen in this step. Click next.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 18.1: Sample Grouping and Significance Tests I

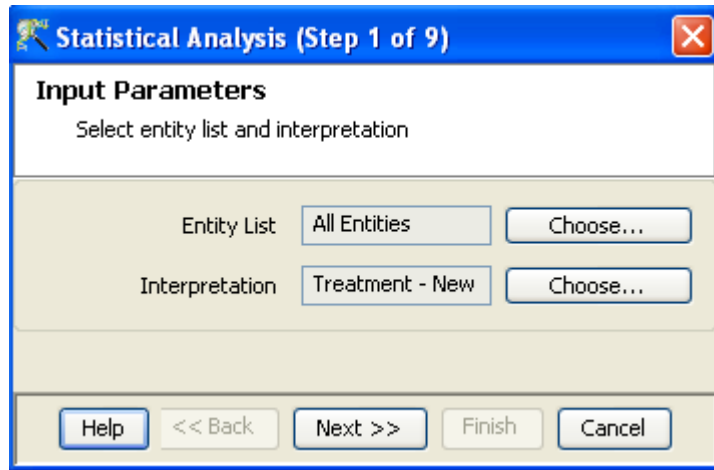


Figure 18.14: Input Parameters

Step 2 of 9: This step allows the user to choose pairing among the groups to be compared, i.e. "a" vs "b" or "b" vs "a". For the kind of experimental design (table above), several tests exist-t-test unpaired, t-test paired, t-test unpaired unequal variance, Mann Whitney unpaired and Mann Whitney paired. Choose the desired test. See Figure [18.15](#)

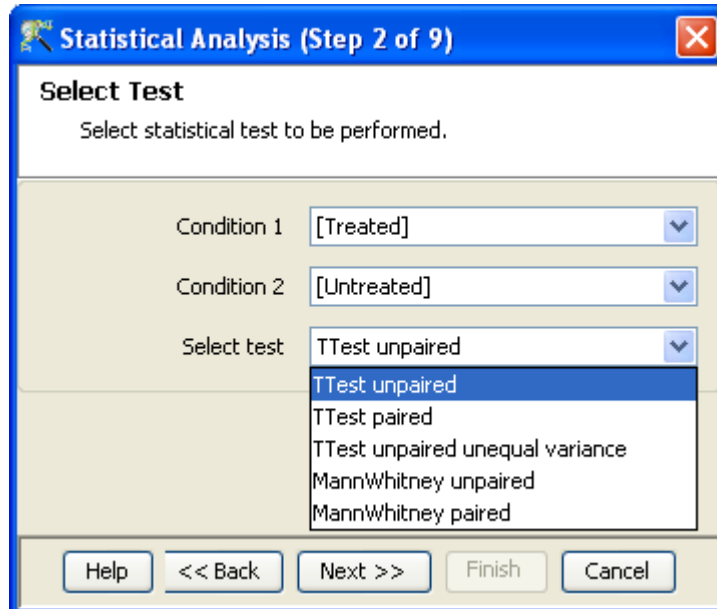


Figure 18.15: Select Test

Steps 3, 4 and 5 of 9: The steps 3 , 4 and 5 are invoked in cases where ANOVA and t-test against zero are to be used. Based upon the experiment design, **GeneSpring GX** goes to the appropriate steps.

Step 6 of 9: p-value computation algorithm and the type of p-value correction to be done are chosen here. When permutative computation is chosen, it is recommended that the user increases the number of permutations till convergence is reached. Once convergence is reached, the p-values of the entities remain the same for n as well as n+x number of permutations. See Figure [18.16](#)

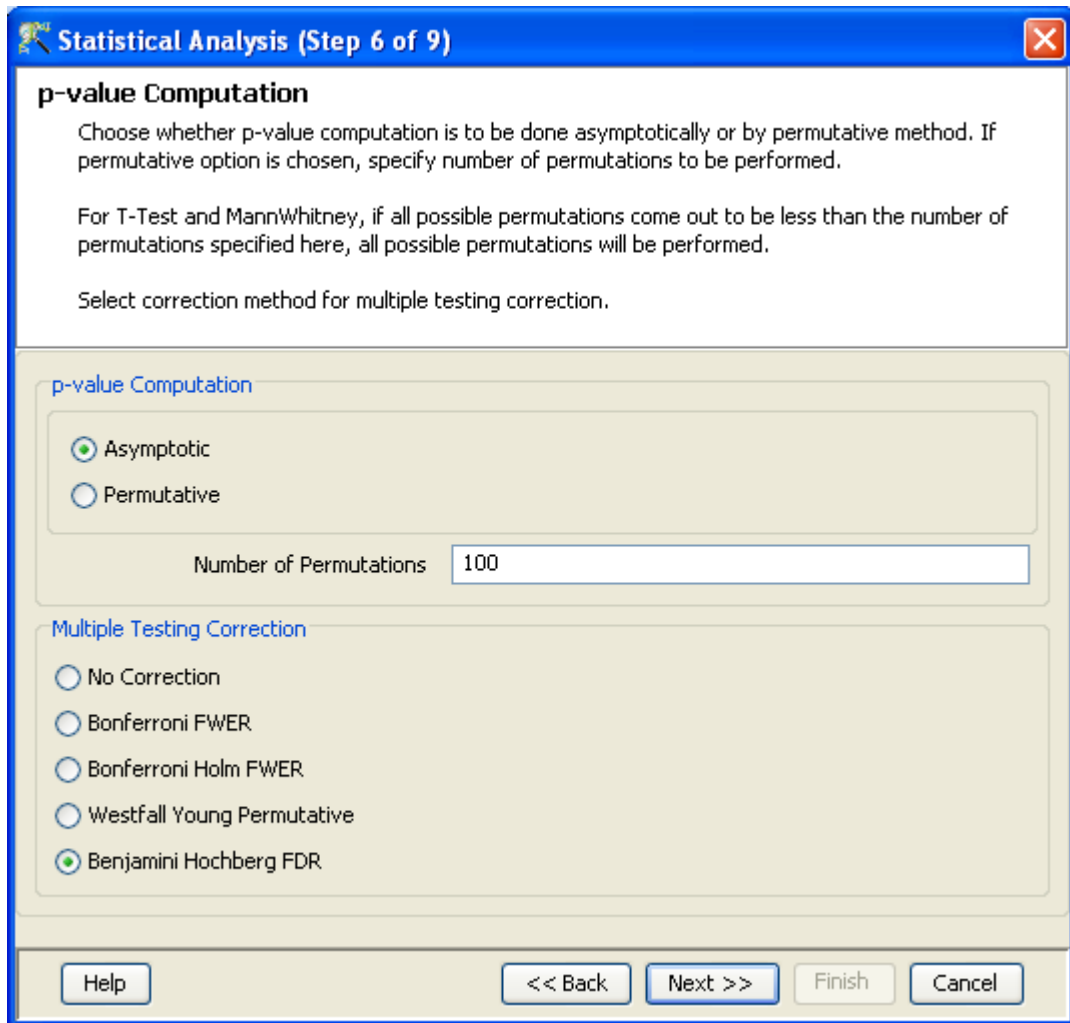


Figure 18.16: p-value Computation

Step 8 of 9: Results of analysis: Upon completion of T-test the results are displayed as three tiled windows.

- A *p-value table* consisting of *Probe Names*, *p-values*, *corrected p-values*, *Fold change (Absolute)* and *regulation*. FC Absolute means that the fold-change reported is absolute. In other words, if an entity is 2-fold up or 2-fold down, it will still be called as 2.0 fold, instead of being called 2.0 fold (for up-regulation) and 0.5 (for down-regulation). Absolute essentially means that there is no directionality associated with the value. Directionality or regulation is indicated separately under the regulation column
- *Differential expression analysis report* mentioning the Test description i.e. the test that has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic* or *Permutative*). Also gives a result summary with different p-value cut-off.
- *Volcano plot* comes up only if there are two groups provided in *Experiment Grouping*. The entities which satisfy the default p-value cutoff 0.05 appear in red colour and the rest appear in grey colour. This plot shows the negative log₁₀ of p-value vs log_(base2.0) of fold change.

Probesets with large fold-change and low p-value are easily identifiable on this view. If no significant entities are found then p-value cut off can be changed using *Rerun Analysis* button. An alternative control group can be chosen from *Rerun Analysis* button. The label at the top of the wizard shows the number of entities satisfying the given p-value.

The views differ based upon the tests performed.

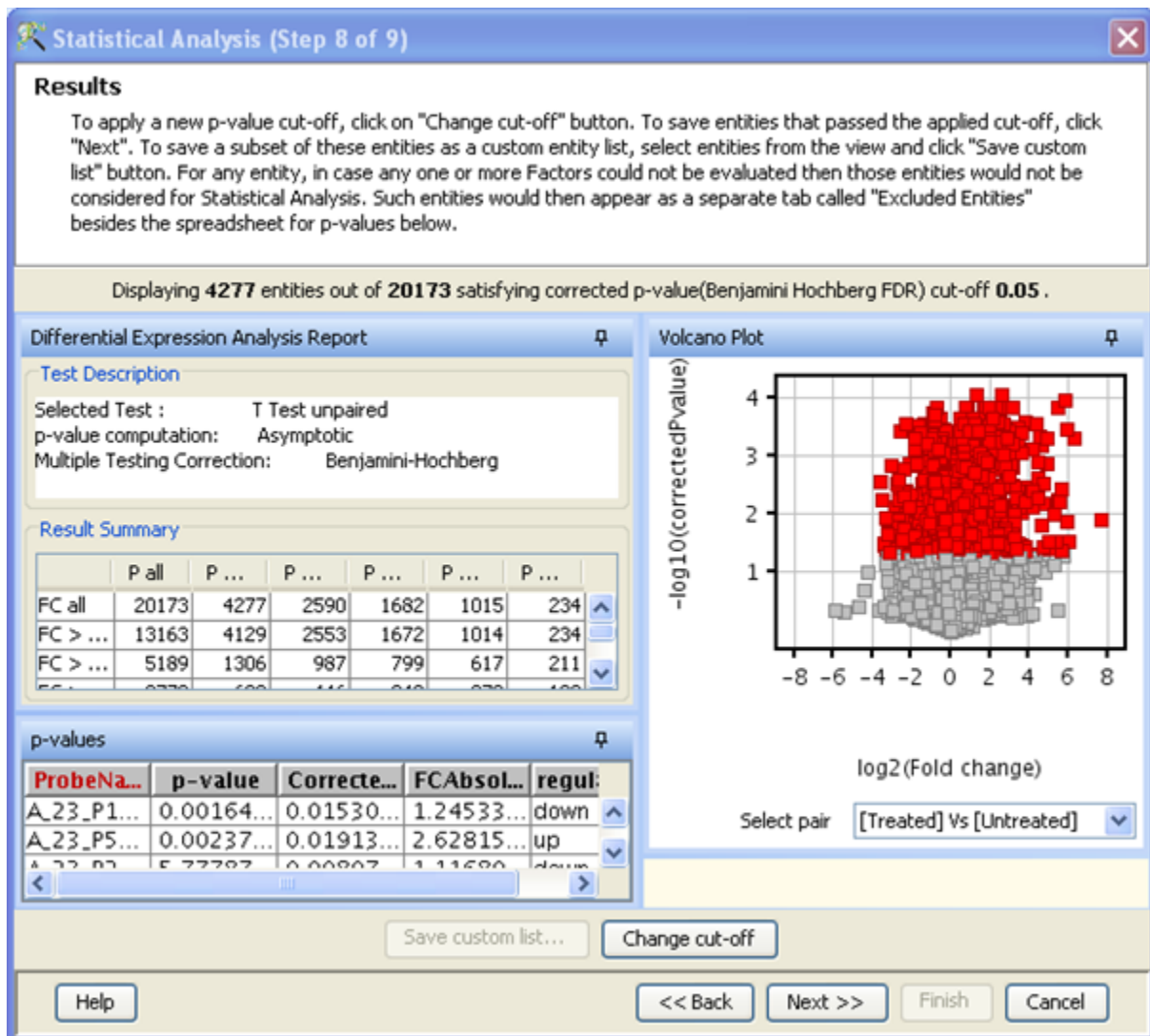


Figure 18.17: Results

Step 9 of 9: The last page shows all the entities passing the p-value cutoff along with their annotations. It also shows the details (regarding Creation date, modification date, owner, number of entities, notes etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to entities which satisfied the cutoff. The name of the entity list will be displayed in the experiment navigator. Annotations can be configured using *Configure Columns* button.

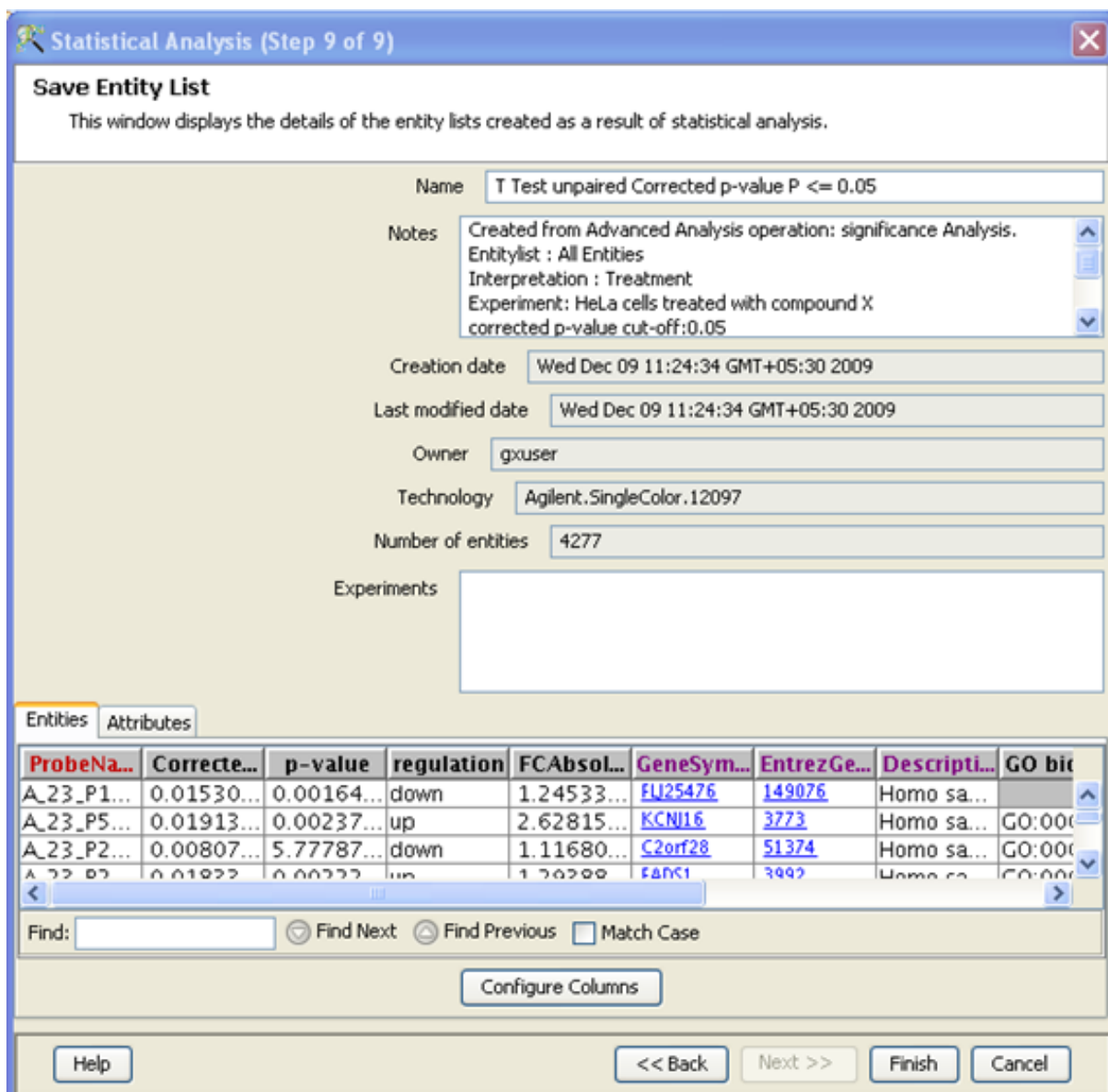


Figure 18.18: Save Entity List

Depending upon the experimental grouping, **GeneSpring GX** performs either T-test or ANOVA. The tables below give information on the type of statistical test performed given any specific experimental grouping:

- **Example Sample Grouping I:** The example outlined in the table *Sample Grouping and Significance Tests I* 18.2, has 2 groups, the Normal and the tumor, with replicates. In such a situation, unpaired t-test will be performed.
- **Example Sample Grouping II:** In this example outlined in table 18.3, only one group, the Tumor, is present. t-test against zero will be performed here.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor
S5	Tumor
S6	Tumor

Table 18.2: Sample Grouping and Significance Tests I

Samples	Grouping
S1	Tumor
S2	Tumor
S3	Tumor
S4	Tumor
S5	Tumor
S6	Tumor

Table 18.3: Sample Grouping and Significance Tests II

- **Example Sample Grouping III:** When 3 groups are present (Normal, tumor1 and Tumor2) and one of the groups (Tumour2 in this case) does not have replicates (shown in table 18.4, statistical analysis cannot be performed. However if the condition Tumor2 is removed from the interpretation (which can be done only in case of *Advanced Analysis*), then an unpaired t-test will be performed.

Samples	Grouping
S1	Normal
S2	Normal
S3	Normal
S4	Tumor1
S5	Tumor1
S6	Tumor2

Table 18.4: Sample Grouping and Significance Tests III

- **Example Sample Grouping IV:** When there are 3 groups within an interpretation as shown in table 18.5, One-way ANOVA will be performed. When ANOVA is run, an additional step, Step 7 of 9 is shown for giving pairing option for 'Fold Change Analysis'. In the results page shown in step 8, Fold change values are reported along with p values. This step is shown in Figure 18.19
- **Example Sample Grouping V:** The table 18.6 shows an example of the tests performed when 2 parameters are present. Note the absence of samples for the condition Normal/50 min and Tumor/10 min. Because of the absence of these samples, no statistical significance tests will be performed.
- **Example Sample Grouping VI:** In this table 18.7, a two-way ANOVA will be performed.
- **Example Sample Grouping VII:** In the example shown in table 18.8, a two-way ANOVA will

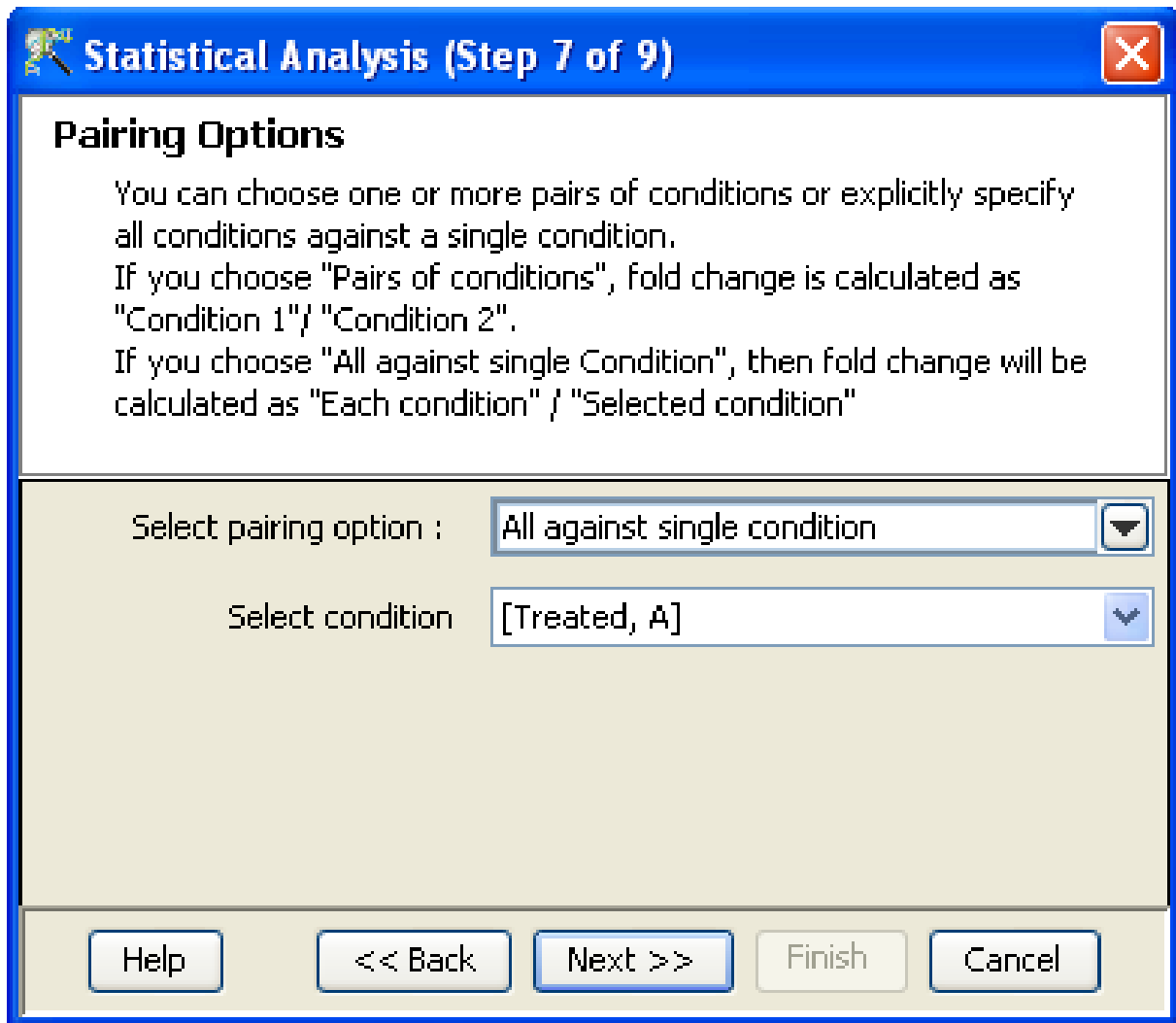


Figure 18.19: Pairing Options

Samples	Grouping
S1	Normal
S2	Normal
S3	Tumor1
S4	Tumor1
S5	Tumor2
S6	Tumor2

Table 18.5: Sample Grouping and Significance Tests IV

be performed and will output a p-value for each parameter, i.e. for Grouping A and Grouping B. However, the p-value for the combined parameters, Grouping A- Grouping B will not be computed. In this particular example, there are 6 conditions (Normal/10min, Normal/30min, Normal/50min, Tumor/10min, Tumor/30min, Tumor/50min), which is the same as the number of samples. The

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	10 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	50 min

Table 18.6: Sample Grouping and Significance Tests V

Samples	Grouping A	
S1	Normal	10 min
S2	Normal	10 min
S3	Normal	50 min
S4	Tumor	50 min
S5	Tumor	50 min
S6	Tumor	10 min

Table 18.7: Sample Grouping and Significance Tests VI

p-value for the combined parameters can be computed only when the number of samples exceed the number of possible groupings.

Samples	Grouping A	Grouping B
S1	Normal	10 min
S2	Normal	30 min
S3	Normal	50 min
S4	Tumour	10 min
S5	Tumour	30 min
S6	Tumour	50 min

Table 18.8: Sample Grouping and Significance Tests VII

- **Example Sample Grouping VIII:** In the example shown in table 18.9, with three parameters, a 3-way ANOVA will be performed.

Note: If a group has only 1 sample, significance analysis is skipped since standard error cannot be calculated. Therefore, at least 2 replicates for a particular group are required for significance analysis to run.

ANOVA: Analysis of variance or ANOVA is chosen as a test of choice under the experimental grouping conditions shown in the Sample Grouping and Significance Tests Tables IV, VI and VII. The results are displayed in the form of four tiled windows:

Samples	Grouping A	Grouping B	Grouping C
S1	Normal	Female	10
S2	Normal	Male	10
S3	Normal	Male	20
S4	Normal	Female	20
S5	Tumor1	Male	10
S6	Tumor1	Female	10
S7	Tumor1	Female	20
S8	Tumor1	Male	20
S9	Tumor2	Female	10
S10	Tumor2	Female	20
S11	Tumor2	Male	10
S12	Tumor2	Male	20

Table 18.9: Sample Grouping and Significance Tests VIII

- A *p-value table* consisting of Probe Names, p-values, corrected p-values and the SS ratio (for 2-way ANOVA). The SS ratio is the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups.
- *Differential expression analysis report* mentioning the Test description as to which test has been used for computing p-values, type of correction used and P-value computation type (*Asymptotic or Permutative*).
- *Venn Diagram* reflects the union and intersection of entities passing the cut-off and appears in case of 2-way ANOVA.

18.3.2 Filter on Volcano Plot

The *Filter on Volcano Plot* link is present under the Analysis options in the Advanced workflow. This link allows the user to filter entities on volcano plots, which are constructed using fold change values and p-values. Volcano plots allow you to visualize the relationship between fold-change (magnitude of change) and statistical significance (which takes both magnitude of change and variability into consideration). Volcano plots are used to visually represent differential expression between two different conditions and can be used in publications to provide a visual summary of p-values and fold-change values.

After selecting the *Filter on Volcano Plot* option, the user has to go through the following steps:

- The **Input Parameters** window allows the selection of an entity list and an interpretation. This is enabled by selecting the *Choose* button which shows all the available entity lists and interpretations (an option to add a new interpretation is also given). The groups present in the interpretation must have replicates for calculating variance.
- The **Select Test** window allows the selection of the t-test as well as the pair between which the test has to be performed. The user has the option of choosing among the following statistical tests: t-test

paired, t-test unpaired, t-test unpaired unequal variance, MannWhitney unpaired and MannWhitney paired. More information on the above tests is available under section [Details of Statistical Tests in GeneSpring GX](#) . The drop boxes 'Condition 1' and 'Condition 2' allow passing the pair of condition for calculating the fold change. Fold change is calculated as the ratio between Condition 1 and Condition 2.

- If the statistical test chosen is either t-test paired or MannWhitney paired, then the **Column re-ordering** window appears. The reordering can be done by selecting a sample in a column and moving it with the help of the 'up' and 'down' arrow buttons on the side.
- Upon completion of column reordering, the **p-value computation** window appears which allows the selection of a correction method for multiple testing correction. The p-value is computed asymptotically.
- If a statistical test other than the paired tests is chosen, then the window that appears allows the user to select either the Asymptotic or the Permutative option for p-value computation in addition to the options present for multiple testing correction methods. More information on the above options is available in the section [Adjusting for Multiple Comparisons](#)
- The next step shows the results upon completion of the statistical test. They are displayed as four tiled windows.
 1. A p-value table consisting of *Probe Names, p-values, corrected p-values, Fold change (Absolute) and regulation*. FC Absolute means that the fold-change reported is absolute. In other words, if an entity is 2-fold up or 2-fold down, it will still be called as 2.0 fold, instead of being called 2.0 fold (for up-regulation) and 0.5 (for down-regulation). Absolute essentially means that there is no directionality associated with the value. Directionality or regulation is indicated separately under the regulation column as either 'Up' or 'Down'.
 2. *Differential expression analysis report* mentions the test description i.e. which test has been used for computing p-values, type of correction used and P-value computation type (Asymptotic or Permutative).
 3. *Result Summary* shows a tabular column with entities satisfying a range of p-values and Fold Change values.
 4. The *Volcano Plot* displays the entities that satisfy the default p-value cut off 0.05 and a fold change value of 2.0 in red colour and the rest appear in grey colour. This plot shows the negative log₁₀ of p-value vs. log (base2.0) of fold change. The prominent black lines in the plot are provided for visualization purposes and represents the p-value and fold change cut offs in their respective log forms. The user can change the default values by selecting the *Change cutoff* button.

The user can also select entities of interest from either the p-value table, result summary or the volcano plot and save them as an entity list by selecting the *Save Custom list* option.
 5. The label at the top of the wizard shows the number of entities satisfying the given p-value and the fold change.
- The **Save Entity List** window shows the details of the entity list that is created as a result of the above analysis. It also shows information regarding Creation date, modification date, owner, number of entities, notes etc. of the entity list. Annotations can be configured using Configure Columns button. Selecting *Finish* results in an entity list being created containing entities which satisfied the cut off. The name of the entity list will be displayed in the experiment navigator.

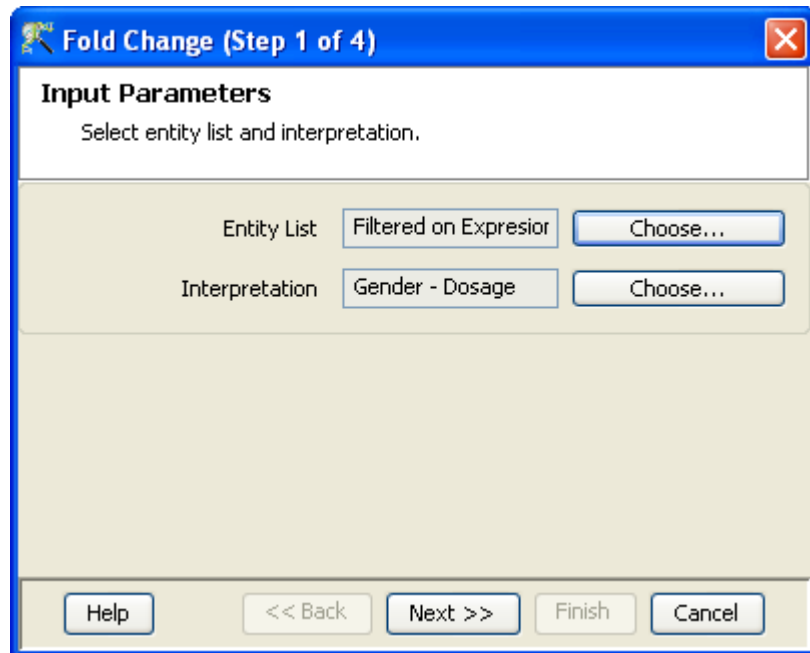


Figure 18.20: Input Parameters

18.3.3 Fold change

Fold Change Analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Fold change is calculated between a condition (Condition 1) and one or more other conditions treated as an aggregate (Condition 2).

Fold change = Condition 1/Condition 2

Fold change gives the absolute ratio of normalized intensities (no log scale) between the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis.

The wizard has following steps:

Step 1 of 4: This step gives an option to select the entity list and interpretation for which fold change is to be evaluated. Note that fold change analysis can be done for 'All samples' interpretation also. Click *Next*.

Step 2 of 4: The second step in the wizard provides the user to select pairing options based on parameters and conditions in the selected interpretation.

Pairing Options:

- Pairs of Conditions : In case of two or more groups, user can evaluate fold change pairwise. The

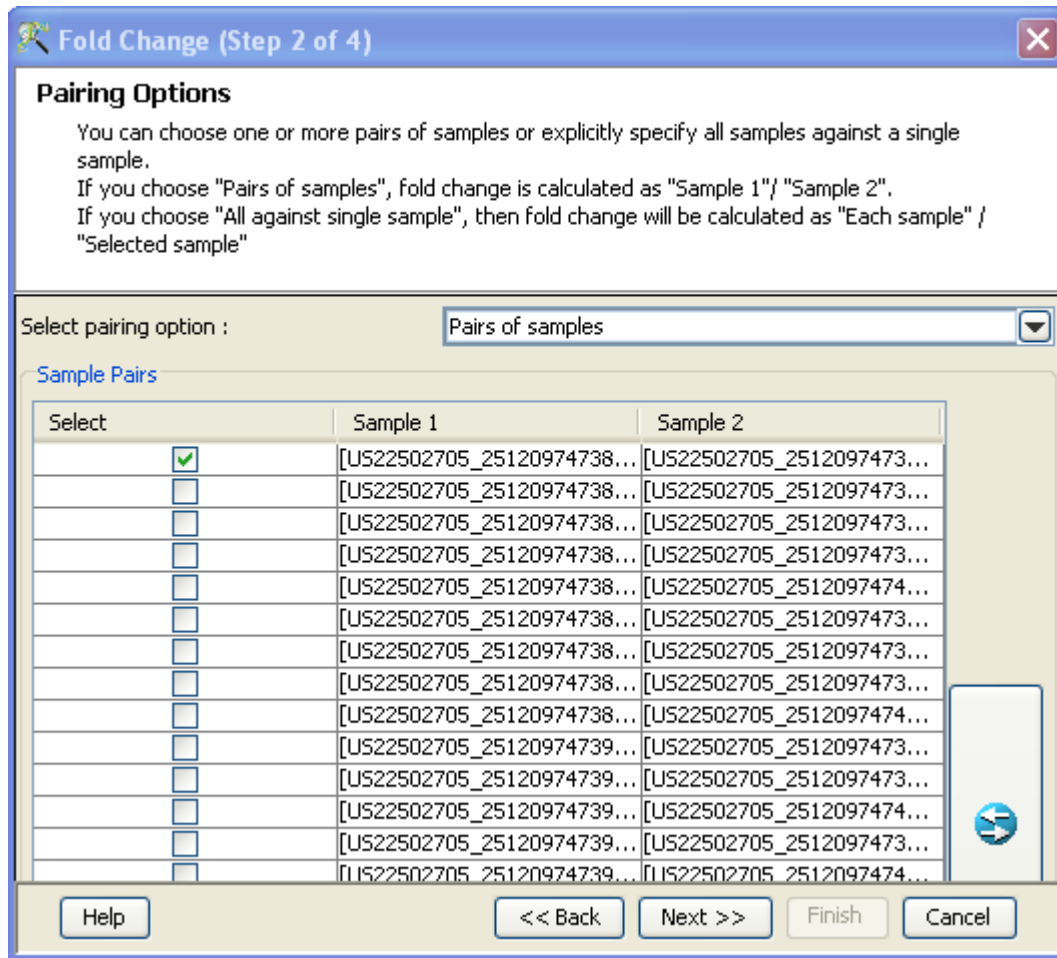


Figure 18.21: Pairing Options

order of conditions can be flipped in case of pairwise conditions using an icon provided in the window.

- **All Against Single Condition:** In this option, each condition (Condition 1) will be paired with the selected condition (Condition 2) . The sample that is to be used as condition 2 needs to be specified in the drop box 'Select Condition'.

Step 3 of 4: This window shows the results in the form of a profile plot and a spreadsheet.

The profile plot shows the up regulated genes in red and down regulated genes in blue color. Irrespective of the pairs chosen for Fold change cutoff analysis, the X-axis of the profile plot displays all the samples. Double click on plot shows the entity inspector giving the annotations corresponding to the selected entity. Selected entities from the plot can be saved using *Save Custom List* button. Fold change cut-off can also be changed in this window.

The columns represented in the spreadsheet are ProbeId, Fold change value and Regulation (up or down) for each fold change analysis. Multiple sets of fold change value and regulation columns would appear in the spreadsheet if 'All against single condition' pairing option was chosen. The regulation

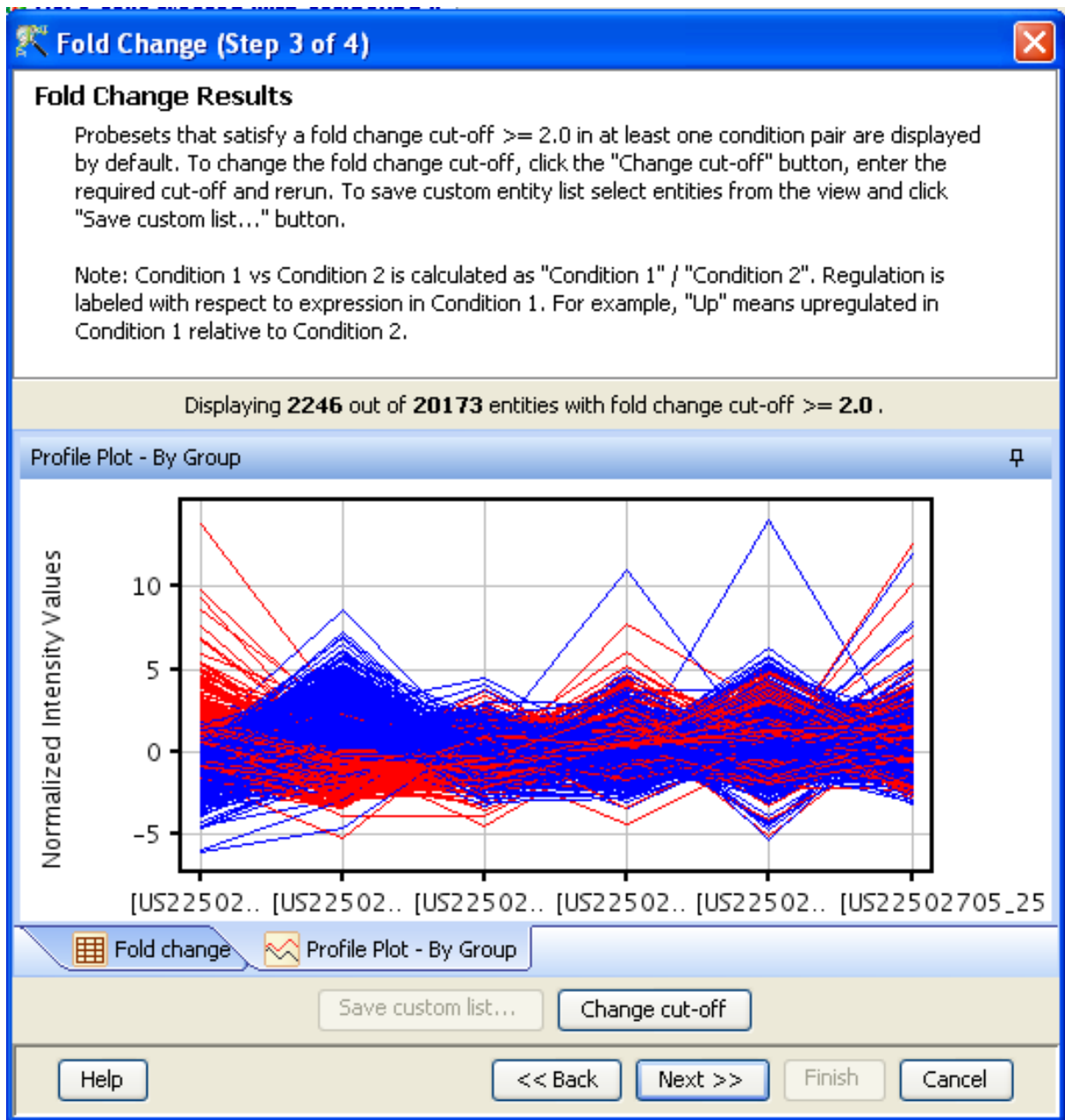


Figure 18.22: Fold Change Results

column can show 'Up' or 'Down' depending on whether Condition 1 has greater or lower intensity values with respect to condition 2. 'Up' means upregulated in Condition 1 relative to Condition 2. The label at the top of wizard shows the number of entities passing the foldchange cut-off. Fold change parameters can be changed by clicking on the change cutoff button and either using the slide bar (goes upto 10) or putting in the desired value and pressing enter. Fold change values cannot be less than 1.

Step 4 of 4: This page shows all the entities passing the fold change cut-off along with their annotations. It also shows the details (regarding Creation date, modification date, owner, number of entities, notes

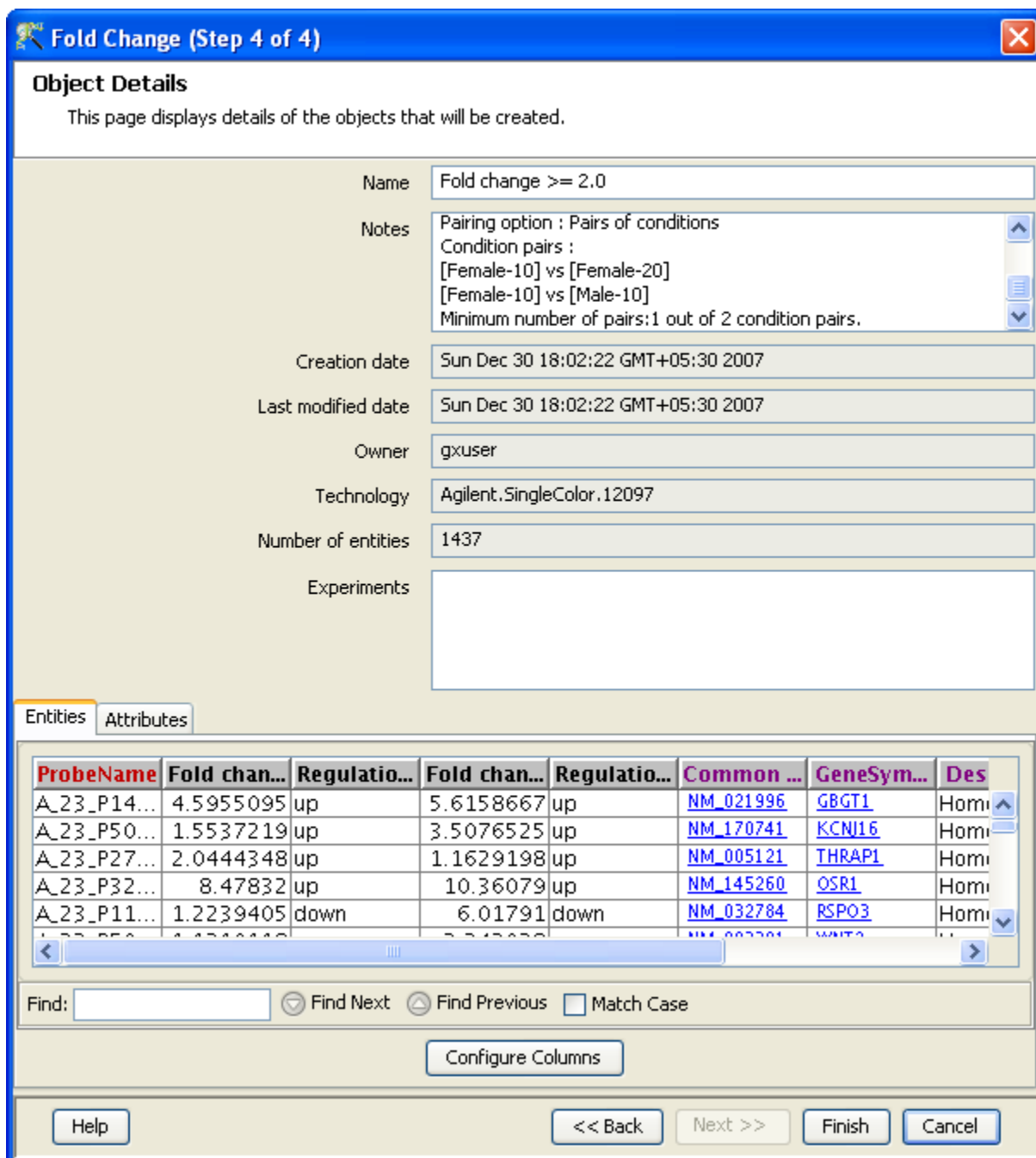


Figure 18.23: Object Details

etc.) of the entity list. Click *Finish* and an entity list will be created corresponding to entities which satisfied the cutoff. Double clicking on an entity in the Profile Plot opens up an *Entity Inspector* giving the annotations corresponding to the selected profile. Additional tabs in the *Entity Inspector* give the raw and the normalized values for that entity. The name of the entity list will be displayed in the experiment navigator. Annotations being displayed here can be configured using *Configure Columns* button.

Note: If multiple conditions are selected for condition one, the fold change for each of the conditions in condition 1 will be calculated.

18.3.4 Clustering

For further details refer to section [Clustering](#)

18.3.5 Find similar entities

The above option allows the user to query a specific entity list or the entire data set to find entities whose expression profile matches that of the entity of interest.

On choosing *Find Similar Entities* under the Analysis section in the workflow, **GeneSpring GX** takes us through the following steps:

Step 1 of 3: This step allows the user to input parameters that are required for the analysis. Entity list and interpretation are selected here. Next, the entity list displaying the profile of our interest has to be selected in the *Choose Query Entity* box. The similarity metric that can be used in the analysis can be viewed by clicking on the dropdown menu. The options that are provided are:

1. **Euclidean:** Calculates the Euclidean distance where the vector elements are the columns. The square root of the sum of the square of the A and the B vectors for each element is calculated and then the distances are scaled between -1 and +1. Result = $(\mathbf{A}-\mathbf{B}).(\mathbf{A}-\mathbf{B})$.
2. **Pearson Correlation:** Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\mathbf{A}.\mathbf{B}/(-\mathbf{A}-\mathbf{B}-)$
3. **Spearman Correlation:** It orders all the elements of vector **a** and uses this order to assign a rank to each element of **a**. It makes a new vector **a'** where the i-th element in **a'** is the rank of \mathbf{a}_i in **a** and then makes a vector **A** from **a'** in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, it makes a vector **B** from **b**. Result = $\mathbf{A}.\mathbf{B}/(-\mathbf{A}-\mathbf{B}-)$. The advantage of using Spearman Correlation is that it reduces the effect of the outliers on the analysis.

Step 2 of 3: This step allows the user to visualize the results of the analysis in the form of a profile plot. The plot displays the mean profile of the entities that have passed the similarity cut-off. The default range for the cutoff is Min-0.95 and Max-1.0. The cutoff can be altered by using the Change Cutoff button provided at the bottom of the wizard. After selecting the profiles in the plot, they can be saved as an entity list by using the option *Save Custom List*.

Step 3 of 3: This step allows the user to save the entity list created as a result of the analysis and also shows the details of the entity list. Option to configure columns that enables the user to add columns of interest from the given list is present. Clicking on *Finish* creates the entity list which can be visualized under the analysis section of the experiment in the project navigator.

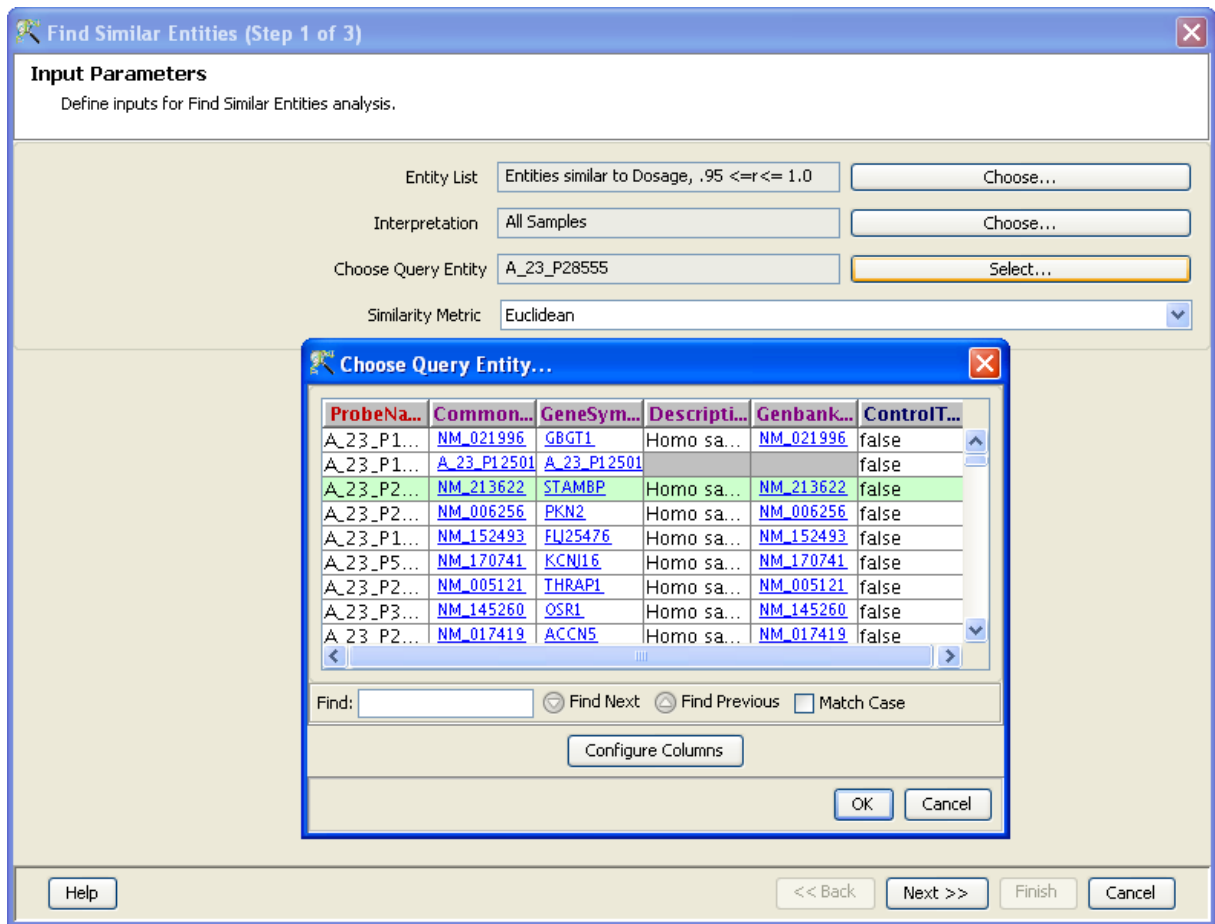


Figure 18.24: Input Parameters

18.3.6 Filter on Parameters

Filter on Parameters calculates the correlation between expression values and parameter values. This filter allows you to find entities that show some correlation with any of the experiment parameters. This filter only works for numerical parameters.

On choosing *Filter on Parameters* under the Analysis section in the workflow, **GeneSpring GX** takes us through the following steps:

Step 1 of 3: This step allows the user to input parameters that are required for the analysis. The entity list and the interpretation are selected here. Also the experiment parameter of our interest has to be selected in the Parameter box. The similarity metric that can be used in the analysis can be viewed by clicking on the dropdown menu. The options that are provided are:

1. **Euclidean:** Calculates the Euclidean distance where the vector elements are the columns. The square root of the sum of the square of the A and the B vectors for each element is calculated and then the distances are scaled between -1 and +1. Result = $(\mathbf{A}-\mathbf{B}) \cdot (\mathbf{A}-\mathbf{B})$.

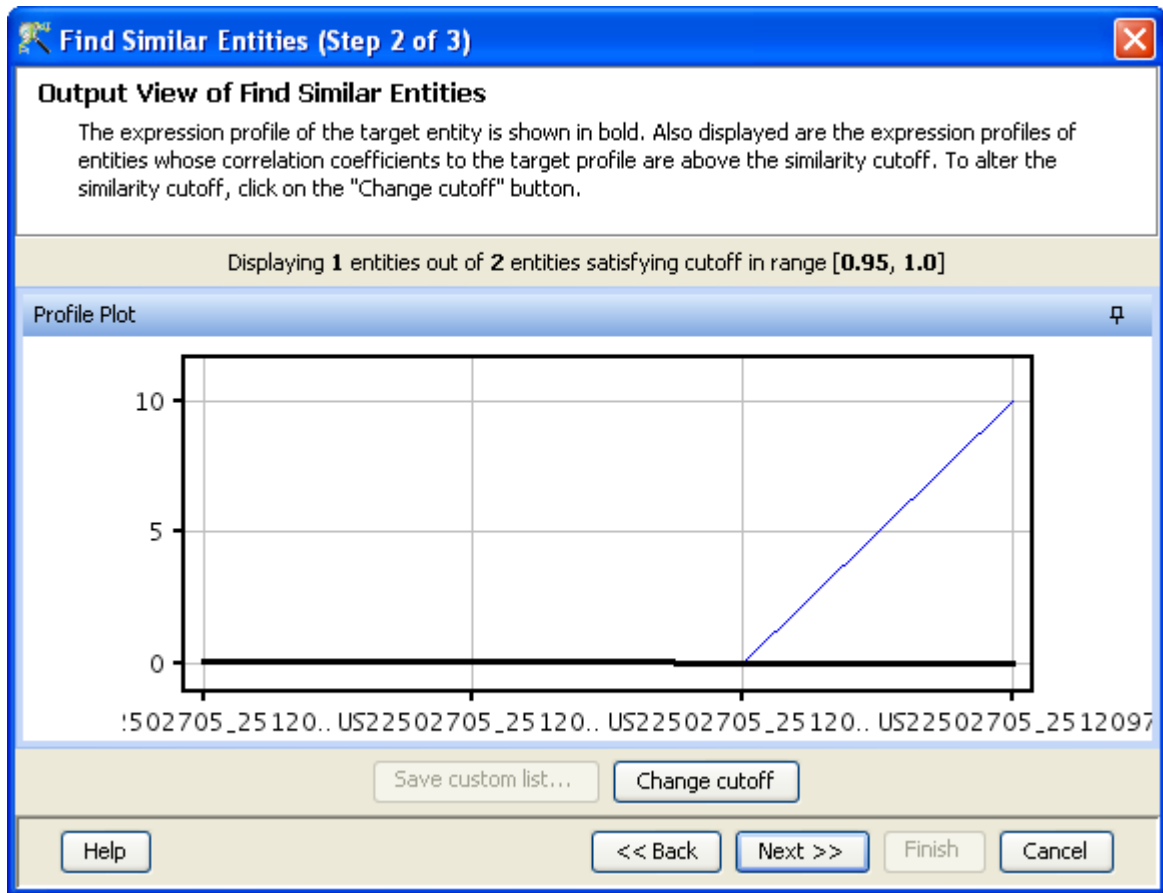


Figure 18.25: Output View of Find Similar Entities

2. **Pearson Correlation:** Calculates the mean of all elements in vector **a**. Then it subtracts that value from each element in **a** and calls the resulting vector **A**. It does the same for **b** to make a vector **B**. Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$
3. Spearman Correlation: It orders all the elements of vector **a** and uses this order to assign a rank to each element of **a**. It makes a new vector **a'** where the *i*-th element in **a'** is the rank of a_i in **a** and then makes a vector **A** from **a'** in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, it makes a vector **B** from **b**. Result = $\mathbf{A} \cdot \mathbf{B} / (\|\mathbf{A}\| \|\mathbf{B}\|)$. The advantage of using Spearman Correlation is that it reduces the effect of the outliers on the analysis.

Step 2 of 3: This step allows the user to visualize the results of the analysis in the form of a profile plot. The plot displays the mean profile of the entities that have passed the similarity cut-off. The default range for the cutoff is Min - 0.95 and Max - 1.0. The cutoff can be altered by using the *Change Cutoff* button provided at the bottom of the wizard. Also after selecting the profiles in the plot, they can be saved as an entity list by using the option *Save Custom List*.

Step 3 of 3: Here, the created entity list and its details as a result of the analysis is displayed. There is also an option to configure columns that enables the user to add columns of interest from the given list. Clicking on *Finish* creates the entity list which can be visualized in the project navigator.

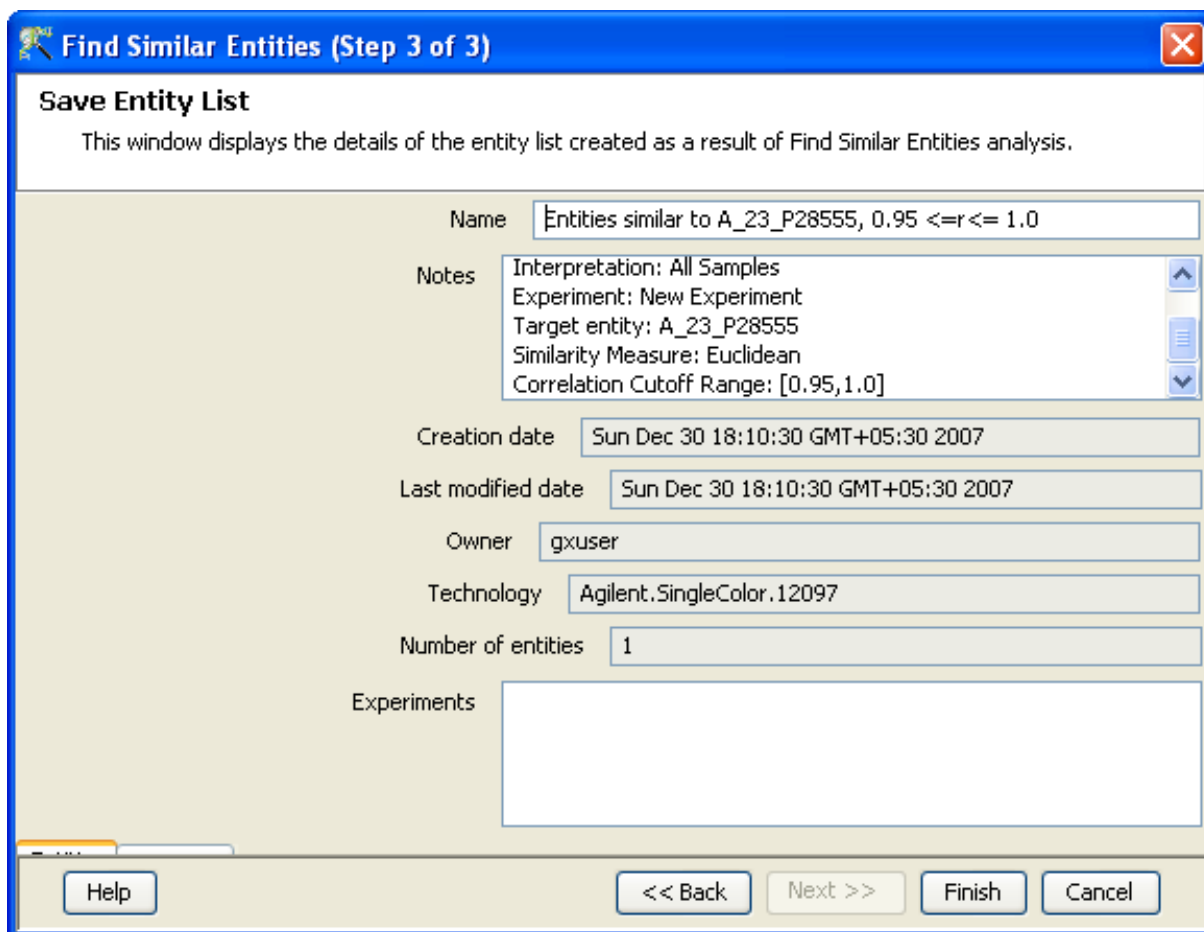


Figure 18.26: Save Entity List

18.3.7 Principal Component Analysis

Viewing Data Separation using Principal Component Analysis:

Imagine trying to visualize the separation between various tumor types given gene expression data for several thousand genes for each sample. There is often sufficient redundancy in these large collection of genes and this fact can be used to some advantage in order to reduce the dimensionality of the input data. Visualizing data in 2 or 3 dimensions is much easier than doing so in higher dimensions and the aim of dimensionality reduction is to effectively reduce the number of dimensions to 2 or 3. There are two ways of doing this - either less important dimensions get dropped or several dimensions get combined to yield a smaller number of dimensions. The Principal Components Analysis (PCA) essentially does the latter by taking linear combinations of dimensions. Each linear combination is in fact an Eigen Vector of the similarity matrix associated with the dataset. These linear combinations (called Principal Axes) are ordered in decreasing order of associated Eigen Value. Typically, two or three of the top few linear combinations in this ordering serve as very good set of dimensions to project and view the data in. These dimensions capture most of the information in the data.

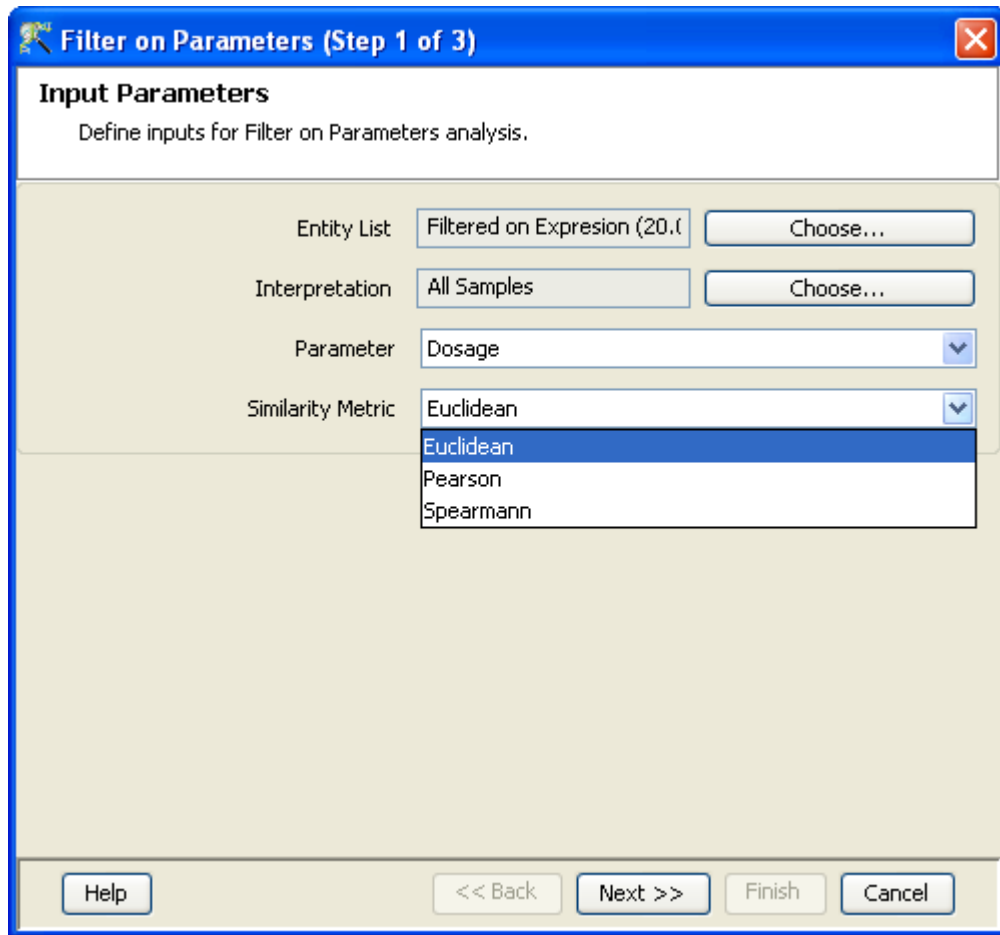


Figure 18.27: Input Parameters

GeneSpring GX supports a fast PCA implementation along with an interactive 2D viewer for the projected points in the smaller dimensional space. It clearly brings out the separation between different groups of rows/columns whenever such separations exist.

The wizard has the following steps:

Step 1 of 3: Entity list and interpretation for the analysis are selected here.

Step 2 of 3: Input parameters for PCA are defined in this step. PCA can either be run on entities (rows) or conditions (columns) of the dataset.

Pruning options for running the PCA can also be defined here. Typically, only the first few eigenvectors (principal components) capture most of the variation in the data. The execution speed of PCA algorithm can be greatly enhanced when only a few eigenvectors are computed as compared to all. The pruning option determines how many eigenvectors are computed eventually. User can explicitly specify the exact number by selecting Number of Principal Components option, or specify that the algorithm compute as many eigenvectors as required to capture the specified Total Percentage Variation in the data.

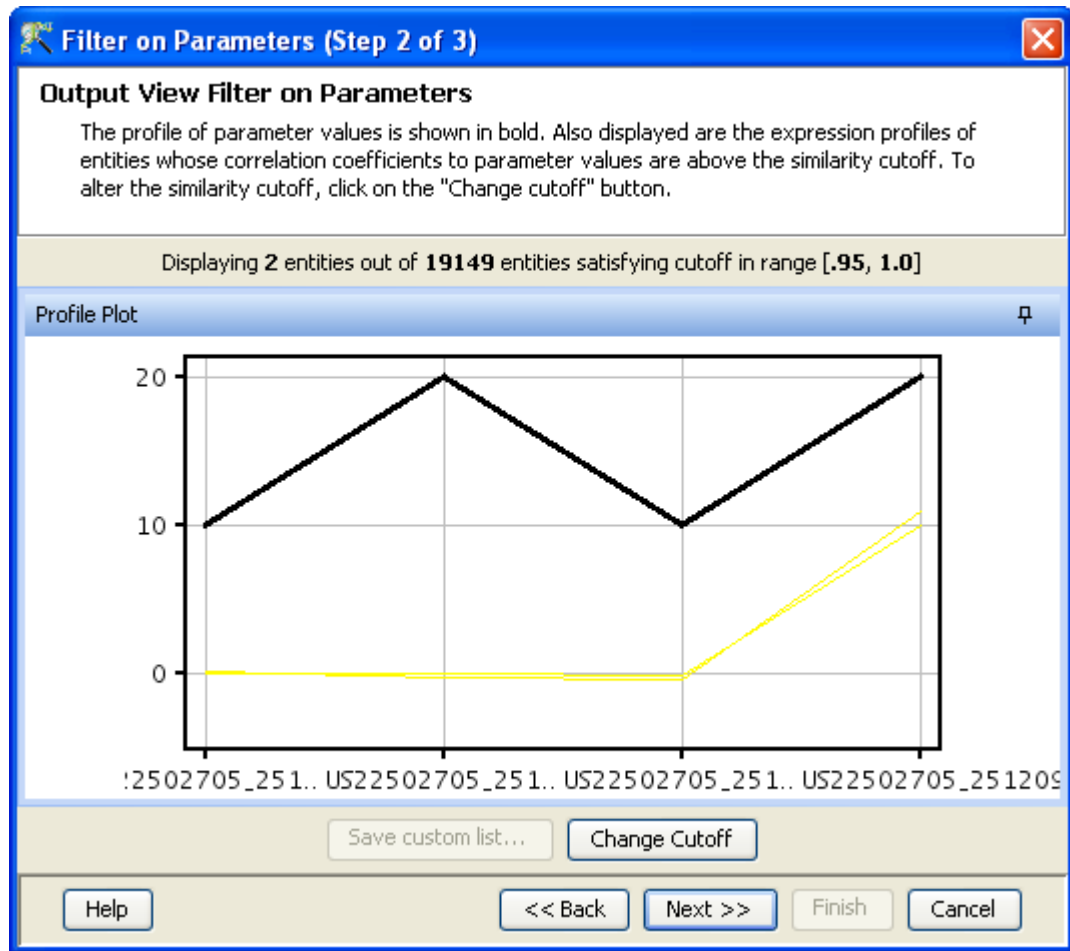


Figure 18.28: Output View of Filter on Parameters

The normalization option to 'mean center' (zero mean) and 'scale' (to unit standard deviation) are enabled by default. Use this if the range of values in the data columns varies widely.

Step 3 of 3: This window shows the Outputs of *Principal Components Analysis*.

The output of PCA is shown in the following four views:

1. **PCA Scores:** This is a scatter plot of data projected along the principal axes (eigenvectors). By default, the first and second PCA components are plotted to begin with, which capture the maximum variation of the data. If the dataset has a class label column, the points are colored with respect to that column, and it is possible to visualize the separation (if any) of classes in the data. Different PCA components can be chosen using the dropdown menu for the X-Axis and Y-Axis; the percentage variation captured by that component is given alongside the component name. Mouse-over on the plot to know more details of the components.
2. **PCA Loadings:** As mentioned earlier, each principal component (or eigenvector) is a linear combination of the selected columns. The relative contribution of each column to an eigenvector is called its loading and is depicted in the PCA Loadings plot. The X-Axis consists of columns, and the Y-Axis denotes the weight contributed to an eigenvector by that column. Each eigen-

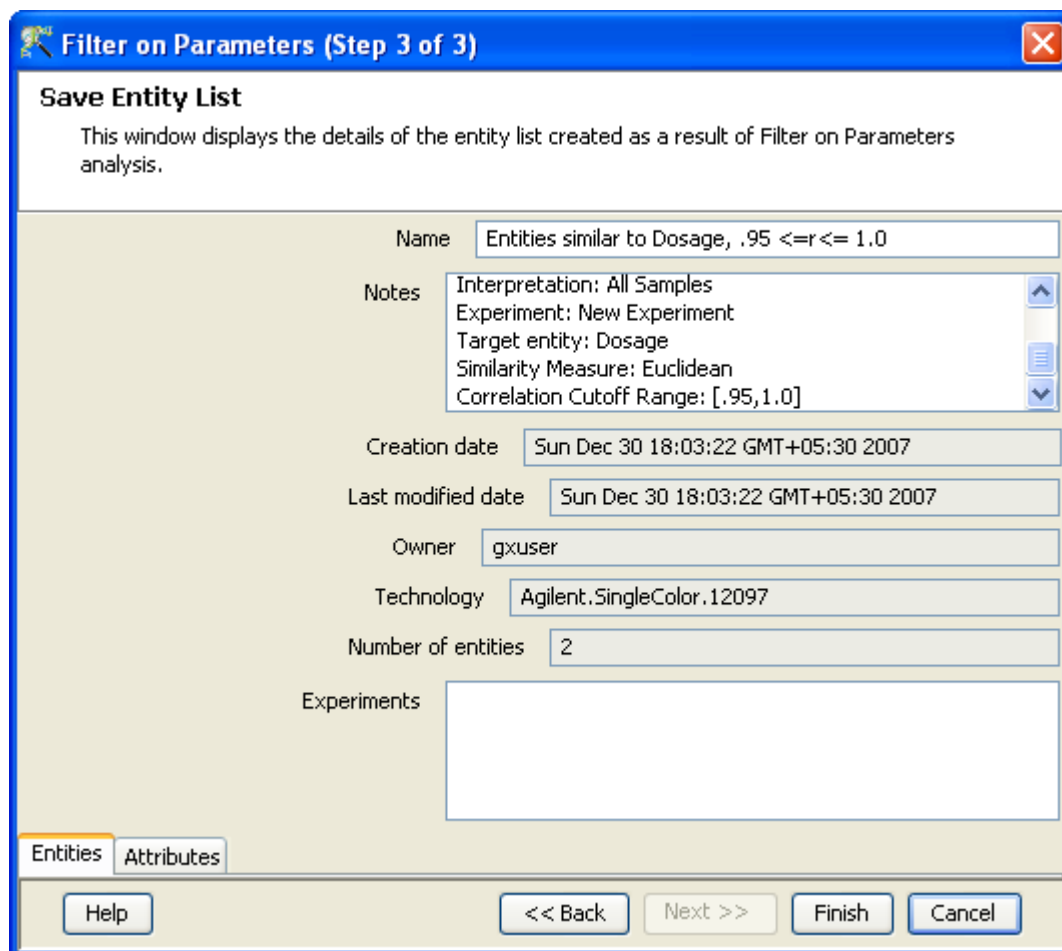


Figure 18.29: Save Entity List

vector is plotted as a profile, and it is possible to visualize whether there is a certain subset of columns which overwhelmingly contribute (large absolute value of weight) to an important eigenvector; this would indicate that those columns are important distinguishing features in the whole data.

3. **Principal Eigen Values:** This is a plot of the Eigen values (Component 1, Component 2, etc.) on X-axis against their respective percentage contribution (Y-axis). The minimum number of principal axes required to capture most of the information in the data can be gauged from this plot. The red line indicates the actual variation captured by each eigen-value, and the blue line indicates the cumulative variation captured by all eigen values up to that point. The minimum value for PCA Eigen values is $(1 * 10^{-3}) / (\text{total number of Principal components})$ and the maximum value is the squareroot of the maximum float value handled by the machine.

4. **Legend:** This shows the legend for the respective active window.

Entities can be selected from the PCA Scores plot and saved using Save custom list button.

Step 4 of 4 This window allows saving the output of *Principal Components Analysis*.

Finish adds a child node titled 'Entity created after PCA' under the experiment.

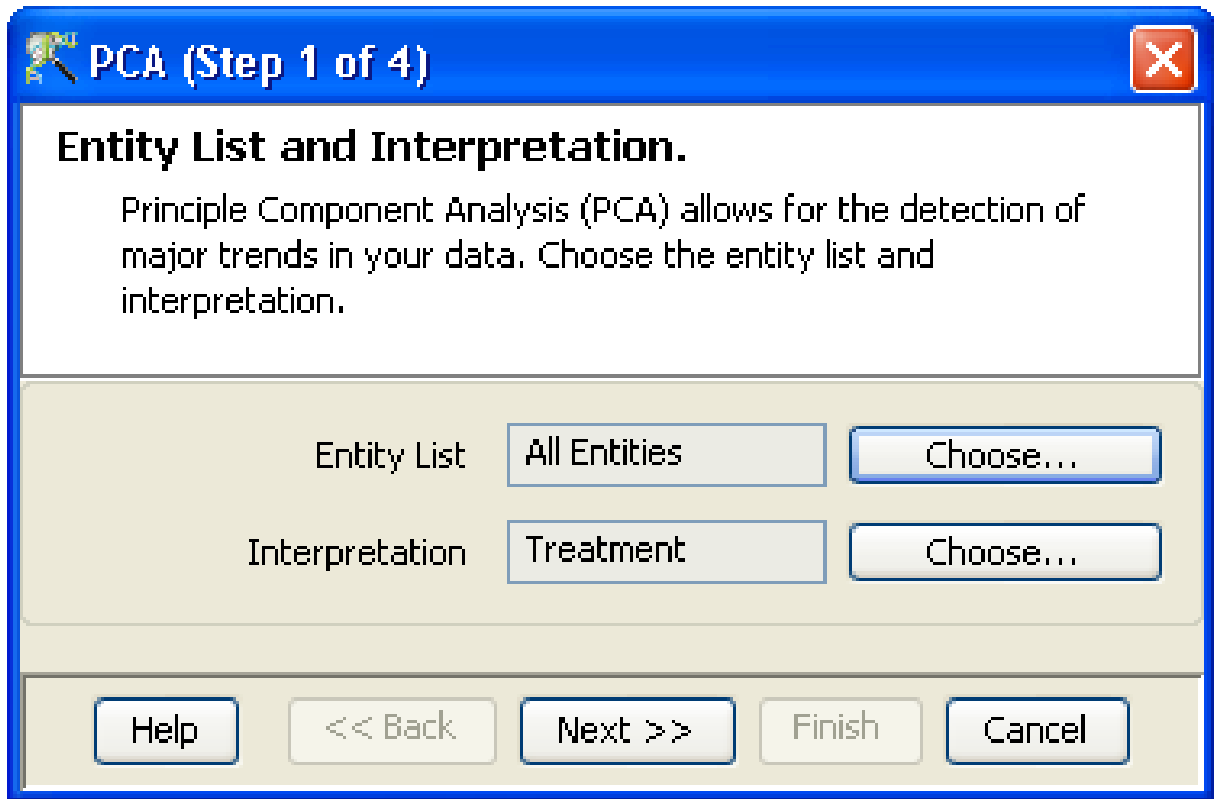


Figure 18.30: Entity List and Interpretation

18.4 Class Prediction

GeneSpring GX has a variety of prediction models that include [Decision Tree \(DT\)](#), [Neural Network \(NN\)](#), [Support Vector Machine \(SVM\)](#), and [Naive Bayesian \(NB\)](#) algorithms. You can build prediction any of these prediction models on the current active experiment that will use the expression values in an entity list to predict the conditions of the interpretation in the current experiment. Once the model has been built satisfactorily, these models can be used to predict the condition given the expression values. Such prediction are being explored for diagnostic purposes from gene expression data.

18.4.1 Build Prediction model

For further details refer to section [Build Prediction Model](#)

18.4.2 Run prediction

For further details refer to section [Run Prediction](#)

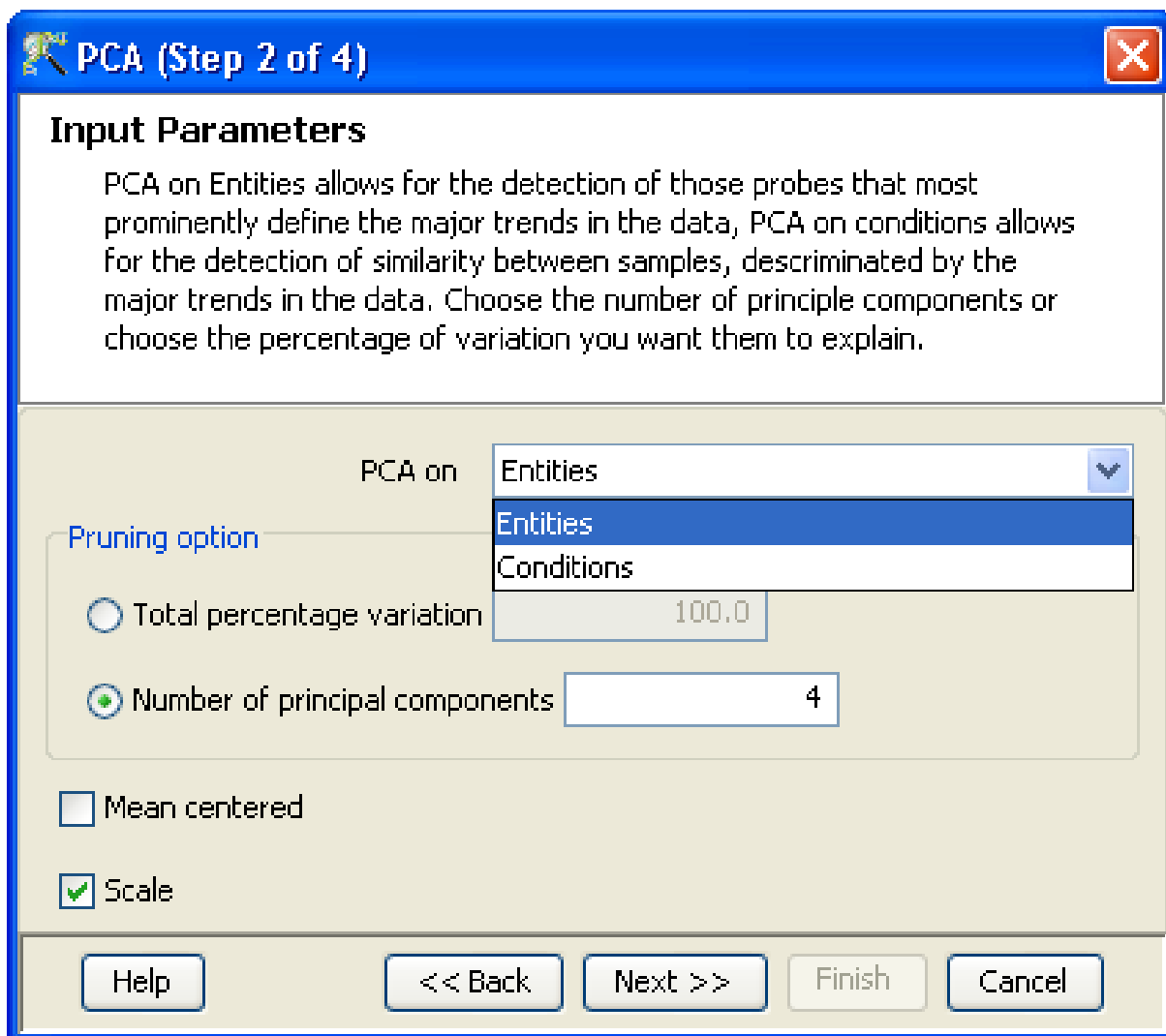


Figure 18.31: Input Parameters

18.5 Results Interpretation

This section contains algorithms that help in the interpretation of the results of statistical analysis. You may have arrived at a set of genes, or an entity list that are significantly expressed in your experiment. **GeneSpring GX** provides algorithms for analysis of your entity list with gene ontology terms. It also provides algorithms for Gene Set Enrichment Analysis or GSEA, which helps you compare your entity list with standard gene sets of known functionality or with your own custom gene sets. In this section, there are also algorithms that help you find entities similar to the chosen entity and to compare the gene lists with metabolic pathways.

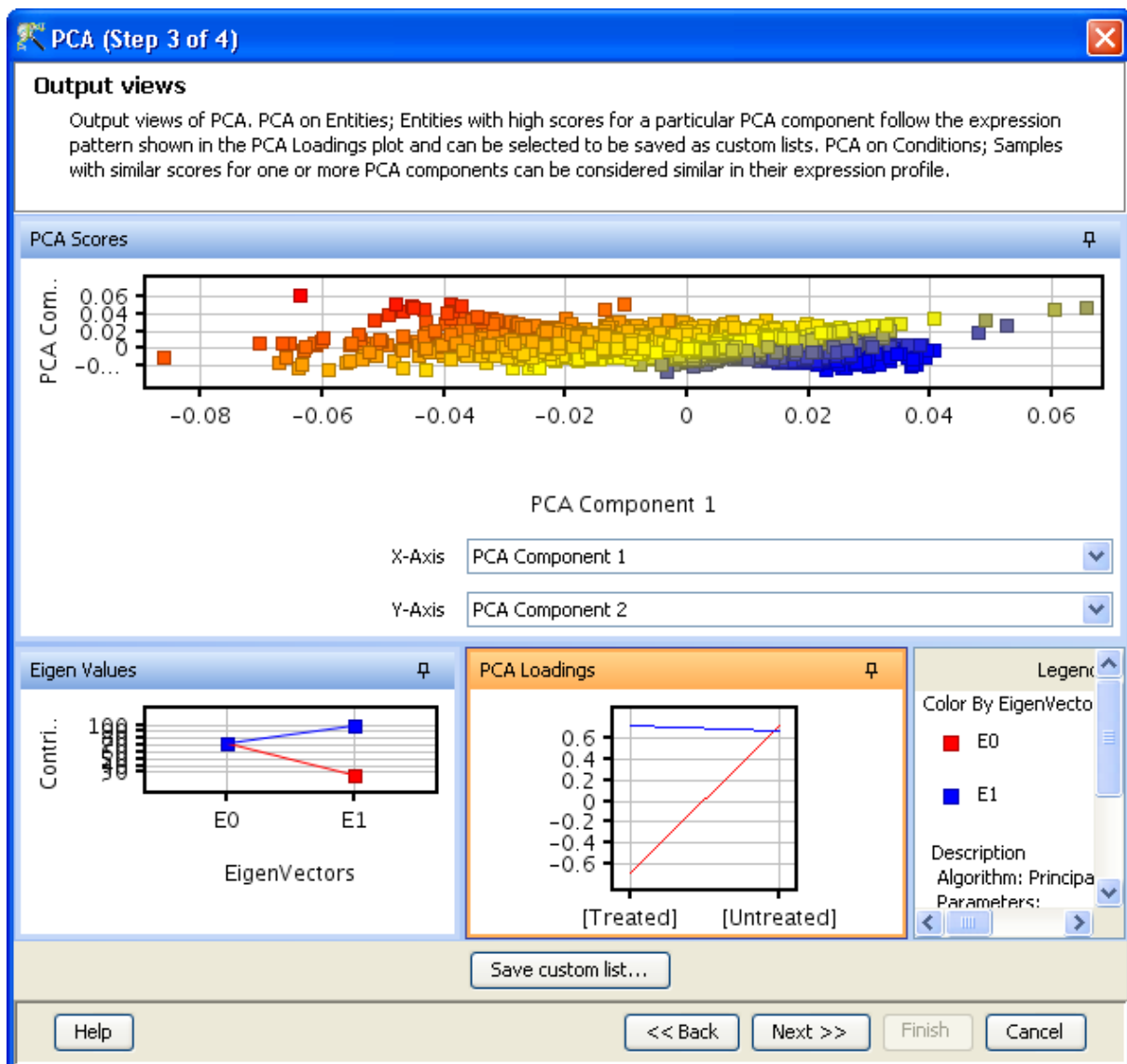


Figure 18.32: Output Views

18.5.1 GO Analysis

Gene Ontology Analysis provides algorithms to explore the Gene Ontology terms associated with the entities in your entity list and calculates enrichment scores for the GO terms associated with your entity list. For a detailed treatment of GO analysis in the refer to the chapter on [GO Analysis](#)

18.5.2 GSEA

Gene set enrichment analysis is discussed in a separate chapter called [Gene Set Enrichment Analysis](#)

18.6 Find Similar Objects

18.6.1 Find Similar Entity lists

Similar entity lists are entity lists that contain a significant number of overlapping entities with the one selected, the significance quantified by P value. Users can select an entity list and start the search by defining the target entity lists and the type of target. The search can be performed even across experiments and projects and on entities belonging to different organisms and technology, provided translation is possible.

A custom search can also be performed where the user can define conditions based on which target entity lists will be pulled out across projects and used for search. Different conditions can be combined using 'OR' and 'AND' feature. Thiswide search area allows user to harness novel information on entities across population.

The wizard to perform this operation has three steps:

1. Step 1 of 3 of *Find Similar Entity Lists*: This step allows the user to first choose the entity list for which similar entity lists are to be found. Then the target entity list and the type of target can be chosen, on which the search will be performed. Under 'Target Entity List', there is an option to choose 'Custom'. This option will allow user to choose target entity lists based on certain conditions, in step 2 of the wizard.
2. Step 2 of 3 of *Find Similar Entity Lists*: This step is shown only if the 'Custom' option has been chosen in Step 1 under Target Entity Lists.

On clicking '*Choose Entity Lists*', a two-step search wizard opens up.

- (a) Step 1 of 2 of '*EntityList Search Wizard*': In the table, choose the required search field, search condition and the value. Multiple searches can be combined by defining 'OR' or 'AND' feature from the drop down 'Combine Search Conditions by'. The conditions can also be defined based on user attributes after checking the item 'Show User Attributes'.

- (b) Step 2 of 2 of *'EntityList Search wizard'*: The results are shown in the form of a table here. Choose those entity lists that need to be searched on and click *Finish* to exit this wizard. The chosen entity lists are now shown as targets in the *'Find Similar Entity Lists'* wizard in step 2.
3. Step 3 of 3: Here the results are presented in the form of two tables. One table showing significant entity lists and the other showing non significant entity lists. The columns in the table list Experiment, query Entity list, Number of entities matching with technology and the query, and p-value. The *p*-value is calculated using the hypergeometric distribution. This equation calculates the probability of overlap corresponding to *k* or more entities between an entity list of *n* entities compared against an entity list of *m* entities when randomly sampled from a universe of *u* genes:

$$(18.1) \frac{1}{\binom{u}{m}} \sum_{i=k}^n \binom{m}{i} \binom{u-m}{n-i}.$$

The p-value cut-off can be changed using *Change Cutoff* button.

To import significant entity list into the experiment, select the entity list and click *Custom Save* button. Click *Finish* and all the similar entity lists will be imported into the active experiment.

18.6.2 Find Similar Pathways

Given an entity list, this functionality enables the user to search and identify pathways whose entities have a significant overlap with the current list. The pathways against which it compares are the BioPax formatted pathways which have been imported and stored.

The *Find Similar Pathways* wizard comprises of 2 steps:

Step 1 of 2: The entity list of interest is specified here.

Step 2 of 2: This step shows 2 windows. The window on the left shows the list of Similar Pathways and the window on the right shows the Non-similar Pathways.

Similar Pathways: This contains the following columns:

1. Pathways: Name of the pathway which passes the p-Value cut-off.
2. Number of Nodes: Total (proteins, small molecules etc) number of nodes in the pathway.
3. Number of Entities: Number of entities from all (genome or array-wide) entities matching with the entities in the pathway.
4. Number of Matching Entities: Number of entities from selected entity list matching with the entities in the pathway.

Non-similar Pathways: This window contains 2 columns, the pathway name and the number of nodes.

Basically one can see this as similar spreadsheet as the Similar Pathways for which **Number of Entities** column has all values zero (i.e. not a single entity from the selected entity list is matching with any of those in that particular pathway).

The level of significance can be modified by selecting the *Change Cutoff* button. Also a significant pathway can be imported into the experiment by selecting the pathway and clicking on the *Custom Save* button. All the similar pathways can be imported into the active experiment by clicking on the *Finish* button. The p-value is calculated in the same way as in the case of *Find Similar Entity Lists* using the equation 18.1

18.7 Utilities

This section contains additional utilities that are useful for data analysis.


18.7.1 Save Current view

Clicking on this option saves the current view before closing the experiment so that the user can revert back to the same view upon reopening the experiment.

18.7.2 Genome Browser

For further details refer to section [Genome Browser](#)

18.7.3 Import Entity List from file

This option allows the user to bring any entity list of interest into the tool. Typically the entity list is a list of probeset IDs or of gene symbols. This functionality is useful when the user wants to view the expression profiles of a select set of genes in any experiment. It can also be used to see the superimposition with pathways or to explore associated GO terms. The entity list should be either in .txt, .csv, .xls, or .tsv formats. Once imported, this list will be added as a child to 'All Entities' list in the Experiment Navigator. The Entity List could be in the form of gene symbols or Probe set IDs or any other annotation which matches with the technology of the active experiment. Import Entity List dialog can be started either from the Utilities section of the workflow or by clicking on the Import Entity List from File  icon on the toolbar. The dialog consists of four fields:

Choose File - This asks the user to specify the path of the file to be imported.

Choose column to match - Here the user has to choose a column that is present in the imported file. This is needed to merge the file with the dataset.

Identifier mark - The column to be imported can be either the probeset ID, Unigene Id or any other annotation. Choose the appropriate mark from the drop-down menu.

Columns to be imported - any other annotation columns to be imported from the Entity List file can be specified here. These additional columns can be brought in only if the Entity List has a Technology Identifier column, otherwise the imported column will be seen as blank.

18.7.4 Import BROAD GSEA Genesets

GSEA can be performed using the 4 genesets which are available from the BROAD Institute's website (<http://www.broad.mit.edu/gsea/>). These genesets can be downloaded and imported into the **GeneSpring GX** to perform GSEA. Clicking on this option allows the user to navigate to the appropriate folder where the genesets are stored and select the set of interest. The files should be present either in .xml or .grp or .gmt formats.

18.7.5 Import BIOPAX pathways

BioPax files required for Pathway analysis can be imported. The imported pathways can then be used to perform *Find Similar Pathways* function. Clicking on this option will allow the user to navigate to the appropriate folder where the files are stored and select the ones of interest. The files should be present in .owl format.

18.7.6 Differential Expression Guided Workflow

Differential Expression Guided Workflow: Clicking on this option launches the Differential Expression Guided Workflow Wizard. This allows the user to switch to *Guided Workflow* from the *Advanced Analysis* when desired.

18.7.7 Filter on Entity List

This utility allows user to filter an Entity list using its annotations and list associated values. The filter can be set by defining a search field, a search condition like 'equals' or 'starts with', and a value for the search field, as applicable. Multiple searches can be combined using OR or AND condition.

The *Filter on Entity List* dialog can be opened from the Utilities section of the workflow.

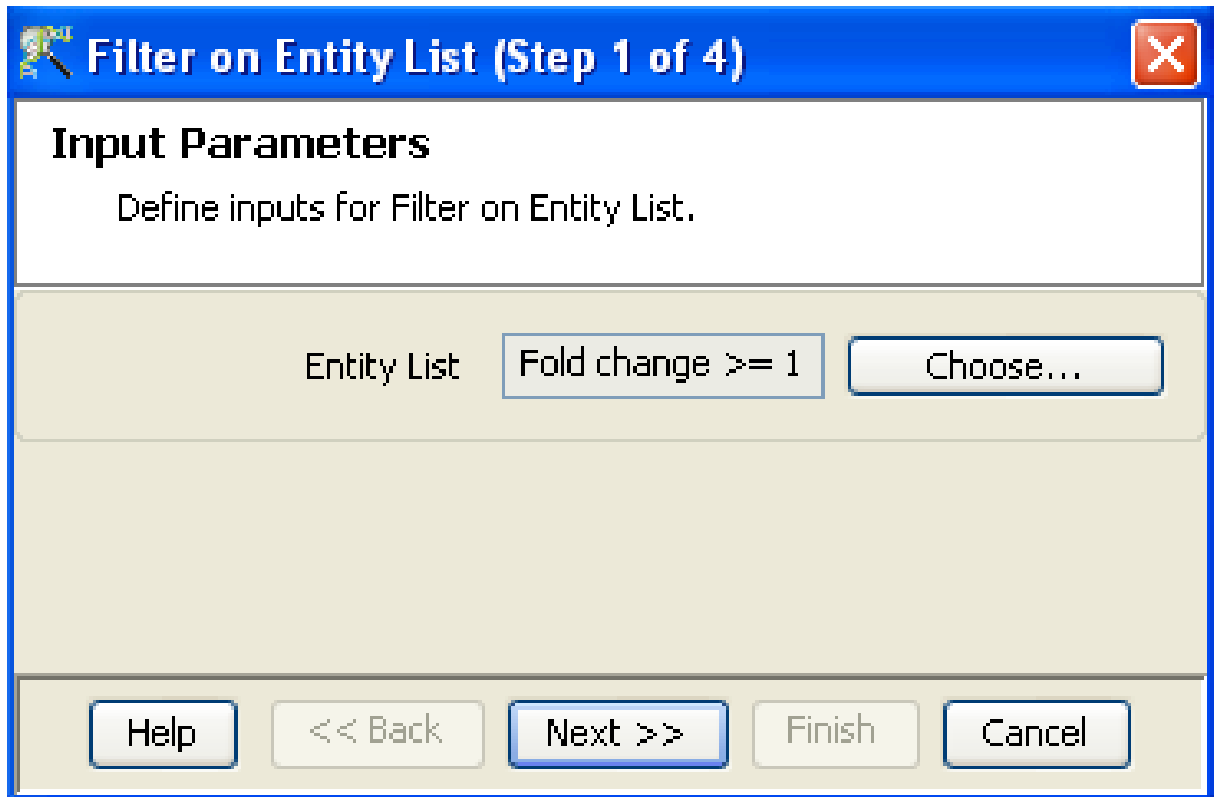


Figure 18.33: Filter on Entity List - Step 1

Filter on Entity List opens a four step wizard.

Step 1 of 4 Allows selection of entity list

Step 2 of 4 Allows defining the filter conditions using three fields: Search field, Condition and Search value.

1. *Search field* Shows all the annotations and list associated values as drop down options
2. *Condition* If the selected search field is a string, the self-explanatory conditions equals, does not equal, starts with, ends with, or includes appear as drop down options. If the selected search field is a numerical field, (for example - Fold change), the options under Condition are their numerical equivalents, =, \neq , \leq , \geq and 'in the range'.
3. *Search value* Allows the desired value (either string or a number, depending on the search field) to be input.

More search conditions can be added or removed using the Add/Remove button. There is also a functionality to combine different search conditions using OR or AND conditions.

Step 3 of 4 The filter results are displayed as a table in this step. Those entities that satisfy the filter conditions are selected by default. All the entities will be selected if the filter conditions are not valid. The selections in the result page can be modified by Ctrl-click.

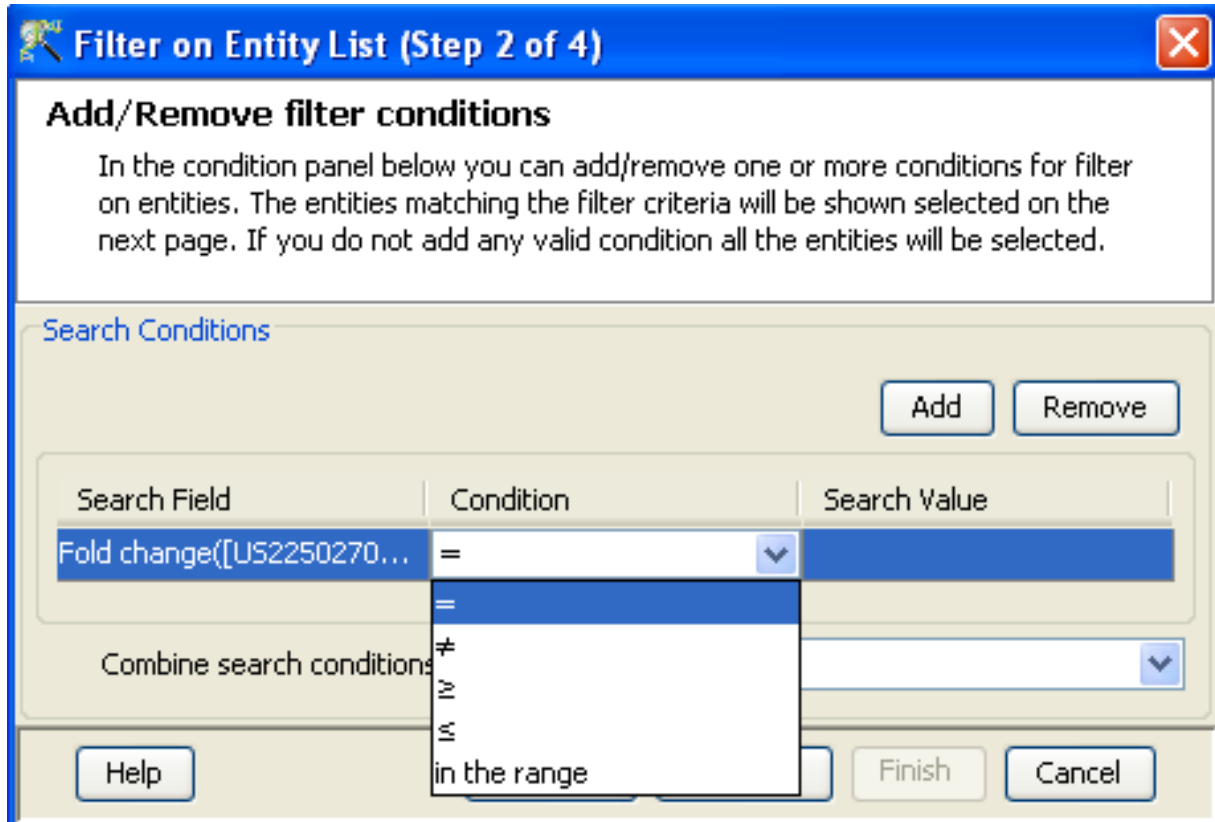


Figure 18.34: Filter on Entity List - Step 2

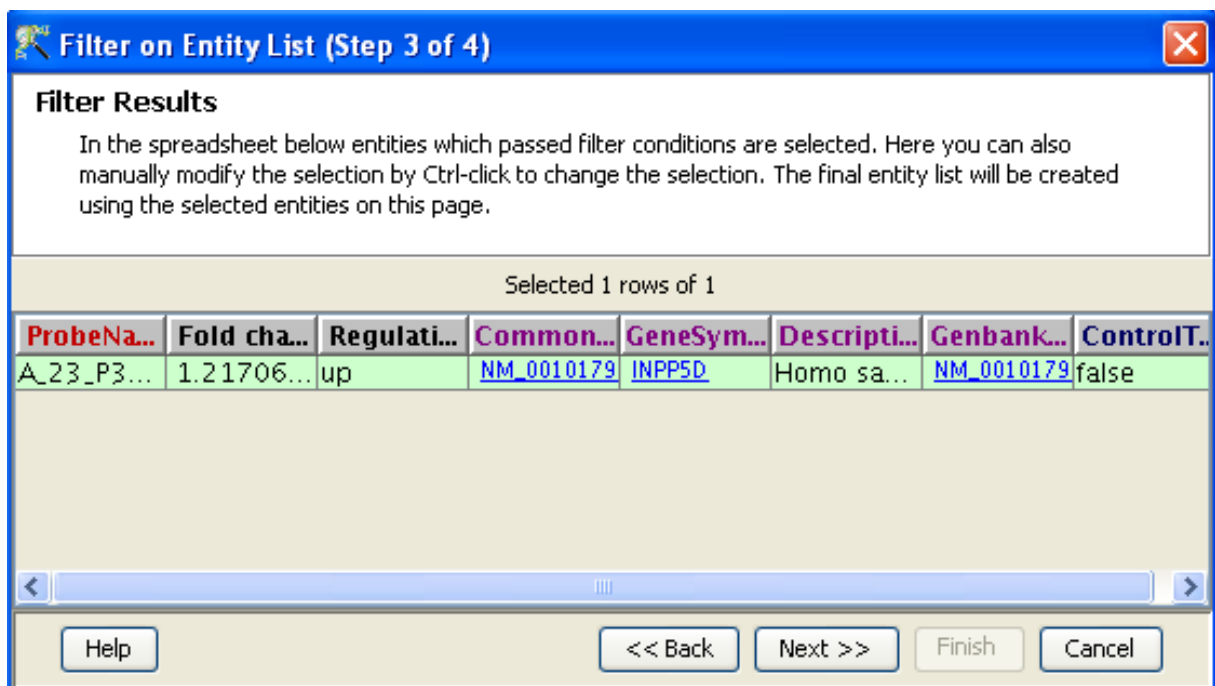


Figure 18.35: Filter on Entity List - Step 3

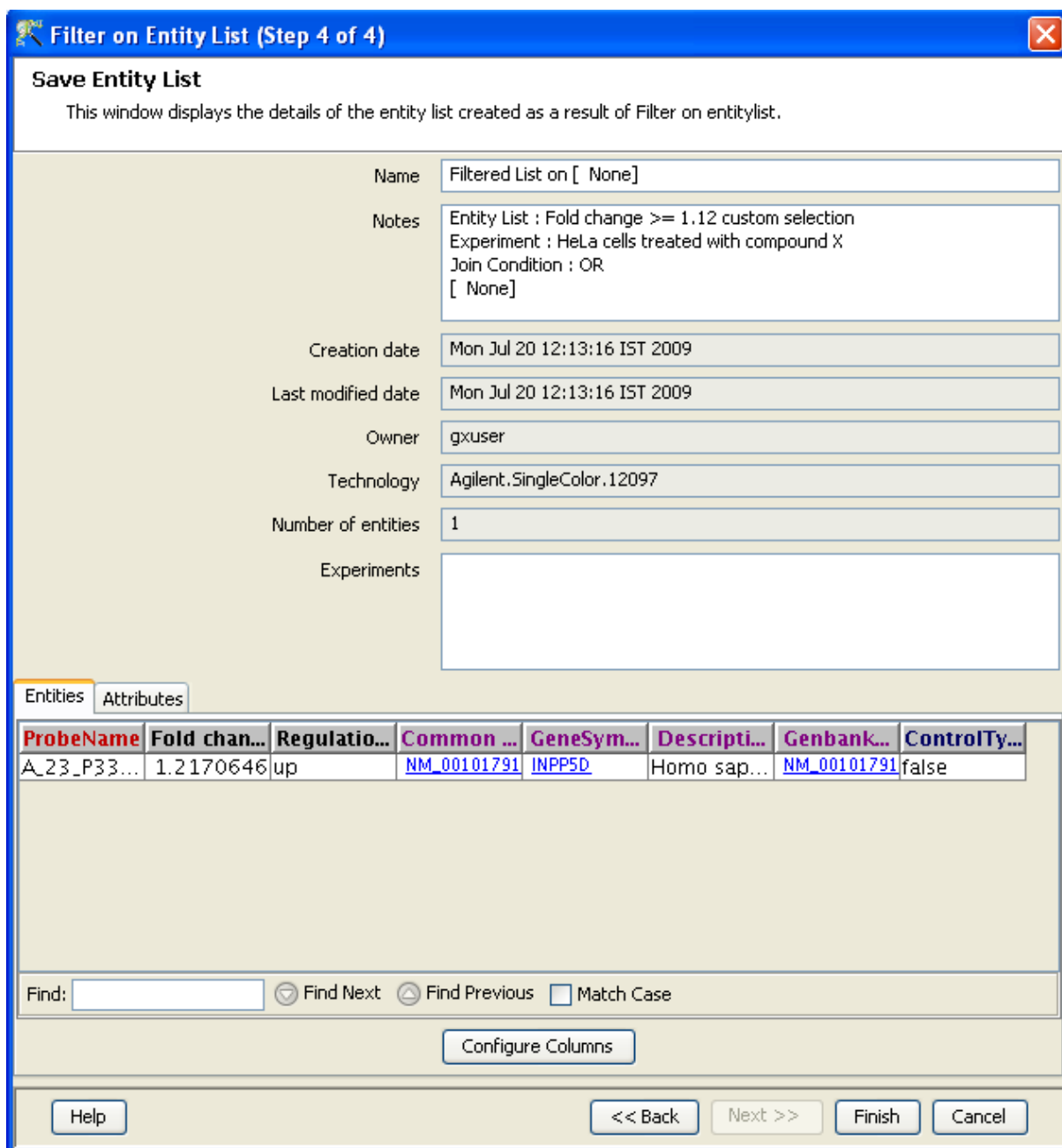


Figure 18.36: Filter on Entity List - Step 4

Step 4 of 4 Allows saving the filtered entity list. Here, the columns in the entity list can be configured before saving. *Finish* will import the filtered entity lists as a child node under the original entity list in the experiment.

Chapter 19

Normalization, Statistical Hypothesis Testing, and Differential Expression Analysis

A brief description of the various normalization methods and statistical tests in **GeneSpring GX** appears below. See [\[24\]](#) for a simple introduction to the statistical tests.

19.1 Threshold

Thresholding is a simple substitution step, wherein all expression values below certain user specified value are made constant, the constant being the specified value. Thresholding is done to remove very small expression values or negative values in the data before taking a log of the data, so that there would not be large negative values or missing values in the log transformed data.

The default in **GeneSpring GX** is to threshold the expression values to 1. If users suspect that bulk of the spots have low expression value then the threshold value should be reduced.

19.2 Normalization Algorithms

A variety of normalization algorithms are used to normalize microarray data consisting of many hybridization samples in an experiment. These are detailed in this section.

19.2.1 Percentile Shift Normalization

Percentile shift normalization is a global normalization, where the location of all the spot intensities in an array are adjusted. This normalization takes each column in an experiment independently, and computes the n^{th} percentile of the expression values for this array, across all spots (where n has a range from 0-100 and $n=50$ is the median). It then subtracts this value from the expression value of each entity.

In **GeneSpring GX**, log transformation is done on the dataset before the normalization and hence the percentile is subtracted from the expression value. Note that for data on linear scale, the expression value would be divided by the percentile.

19.2.2 Scale

This option helps the user in overcoming the inter array differences. **GeneSpring GX** provides scaling option to Median or Mean of control samples/all samples.

If scaling to median is chosen, the normalization method calculates the median of samples (either control samples or all the samples depending on the user specification) individually and then calculates the median of those medians. For example, if S1, S2, S3 and S4 are the samples and the option chosen is to scale the values to the median of all samples, then the median of S1, S2, S3 and S4 are calculated as, say M1, M2, M3 and M4. The next step is to calculate the median (M) of M1, M2, M3 and M4.

A scaling factor is then calculated by subtracting the individual medians (M1, M2, M3 and M4) from the Median of medians (M).

Scaling Factor = $M - M1$; $M - M2$; $M - M3$; $M - M4$ and so on

If the scaling to mean option is chosen, the procedure explained above remains same, but with the mean calculated in place of medians.

This scaling factor is then added to every intensity value on the array. Note that for data in linear scale, the intensity value would be multiplied by the scaling factor, instead of being added.

19.2.3 Quantile Normalization

Quantile normalization is a method of normalization which make the distribution of expression values of all samples in an experiment the same. Thus after normalization, all statistical parameters of the sample,

ie., mean, median and percentiles of all samples will be identical. Quantile normalization works quite well at reducing variance between arrays.

Quantile normalization is performed by the following steps:

- The expression values of each sample is sorted in the ascending order and placed next to each other.
- Each column is sorted in ascending order. The mean of the sorted order across all samples is taken. Thus each row in this sorted matrix has value equal to the previous mean.
- The modified matrix as obtained in the previous step is rearranged to have the same ordering as the input matrix.

Quantile normalization takes care of missing values in the dataset.

19.2.4 Normalize to control genes

This option allows the user to normalize using control genes which can be any of the genes in the array (Rank invariant genes are usually recommended). This option is usually exercised in the case of arrays populated with only specific genes of interest or arrays having less than 1000 spots. It is not advisable to use this normalization if the control genes vary across the samples.

This normalization takes each sample in an experiment independently. It calculates the median of the control genes in each sample and this value is subtracted from all the genes in the sample.

19.2.5 Normalize to External Value

This option is to enable the user to scale the intensity value for each of the sample. Provided a scaling factor for each of the sample, the algorithm subtracts the scaling factor from the signal intensity value, in case of data in log scale. For data in linear scale, the signal intensity value is divided by the scaling factor.

If *Normalization to External Value* is chosen, **GeneSpring GX** will bring up a table listing all samples and a default scaling factor of '1.0'. Users can change this value by using the 'Assign Value' button at the bottom, after highlighting the sample in the table; multiple samples can be chosen simultaneously to assign a value.

19.2.6 Lowess Normalization

In two-color experiments, where two fluorescent dyes (red and green) have been used, intensity-dependent variation in dye bias may introduce spurious variations in the collected data. Lowess normalization merges two-color data, applying a smoothing adjustment that removes such variation.

Lowess normalization characteristics are the following:

- Lowess normalization may be applied to a two-color array expression dataset.
- All samples in the dataset are corrected independently.
- Lowess normalization can be applied to complete or partial datasets. It can be performed independently on each block or portion of the array, or on the whole array.

Lowess regression, or locally weighted least squares regression, is a technique for fitting a smoothing curve to a dataset. The degree of smoothing is determined by the window width parameter. A larger window width results in a smoother curve, a smaller window results in more local variation.

The method involves the following steps:

- Determine the smoothing windows as the percentage of the total number of points or expression values to be considered.
- For the central point in the smoothing window, compute a locally weighted least square regression. Thus points closer to the central point will be given a higher weight and points away from the central point will be given lower weight in the regression. Use this as the value for the central point.
- Move the smoothing window by one point and compute the locally weighted least square regression value for the next central point.
- Repeat this and compute a Lowess normalized expression value for each point of entity in the sample.

The default smoothing parameter for Lowess normalization is 0.2. A sliding window of length 20% of the total number of spots in a grid is used to perform weighted linear regression. Twenty percent of the expression values of all the entities are used to run the locally weighted least square regression. In case the number of spots in a grid are too few (< 250), then a sliding window of length 50 is used to calculate the mean instead of regression.

GeneSpring GX supports Lowess normalization for the whole array or block by block (sub-grid) in all two-color experiments.

19.3 Details of Statistical Tests in GeneSpring GX

19.3.1 The Unpaired t -Test for Two Groups

The standard test that is performed in such situations is the so called t -test, which measures the following t -statistic for each gene g (see, e.g., [24]):

$$t_g = \frac{m_1 - m_2}{s_{m_1 - m_2}}$$

where $s_{m_1 - m_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ is the unbiased pooled variance estimate.

Here, m_1, m_2 are the mean expression values for gene g within groups 1 and 2, respectively, s_1, s_2 are the corresponding standard deviations, and n_1, n_2 are the number of experiments in the two groups. Qualitatively, this t -statistic has a high absolute value for a gene if the means within the two sets of replicates are very different and if each set of replicates has small standard deviation. Thus, the higher the t -statistic is in absolute value, the greater the confidence with which this gene can be declared as being differentially expressed. Note that this is a more sophisticated measure than the commonly used fold-change measure (which would just be $m_1 - m_2$ on the log-scale) in that it looks for a large fold-change in conjunction with small variances in each group. The power of this statistic in differentiating between true differential expression and differential expression due to random effects increases as the numbers n_1 and n_2 increase.

19.3.2 The t -Test against 0 for a Single Group

This is performed on one group using the formula

$$t_g = \frac{m_1}{\sqrt{s_1^2/n_1}}$$

19.3.3 The Paired t -Test for Two Groups

The paired t -test is done in two steps. Let $a_1 \dots a_n$ be the values for gene g in the first group and $b_1 \dots b_n$ be the values for gene g in the second group.

- First, the paired items in the two groups are subtracted, i.e., $a_i - b_i$ is computed for all i .
- A t -test against 0 is performed on this single group of $a_i - b_i$ values.

19.3.4 The Unpaired Unequal Variance t -Test (Welch t -test) for Two Groups

The standard t -test assumes that the variance of the two groups under comparison. Welch t -test is applicable when the variance are significantly different. Welch's t -test defines the statistic t by the following formula:

$$t_g = \frac{m_1 - m_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Here, m_1, m_2 are the mean expression values for gene g within groups 1 and 2, respectively, s_1, s_2 are the corresponding standard deviations, and n_1, n_2 are the number of experiments in the two groups. The degrees of freedom associated with this variance estimate is approximated using the Welch-Satterthwaite equation:

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{\frac{s_1^4}{n_1^2} + \frac{s_2^4}{n_2^2}}{df_1} + \frac{\frac{s_2^4}{n_2^2}}{df_2}}$$

19.3.5 The Unpaired Mann-Whitney Test

The t -Test assumes that the gene expression values within groups 1 and 2 are independently and randomly drawn from the source population **and** obey a normal distribution. If the latter assumption may not be reasonably supposed, the preferred test is the non-parametric Mann-Whitney test, sometimes referred to as the Wilcoxon Rank-Sum test. It only assumes that the data within a sample are obtained from the same distribution but requires no knowledge of that distribution. The test combines the raw data from the two samples of size n_1 and n_2 respectively into a single sample of size $n = n_1 + n_2$. It then sorts the data and provides ranks based on the sorted values. Ties are resolved by giving averaged values for ranks. The data thus ranked is returned to the original sample group 1 or 2. All further manipulations of data are now performed on the rank values rather than the raw data values. The probability of erroneously concluding differential expression is dictated by the distribution of T_i , the sum of ranks for group i , $i = 1, 2$. This distribution can be shown to be normal mean $m_i = n_i(\frac{n+1}{2})$ and standard deviation $\sigma_1 = \sigma_2 = \sigma$, where σ is the standard deviation of the combined sample set.

19.3.6 The Paired Mann-Whitney Test

The samples being paired, the test requires that the sample size of groups 1 and 2 be equal, i.e., $n_1 = n_2$. The absolute value of the difference between the paired samples is computed and then ranked in increasing order, apportioning tied ranks when necessary. The statistic T , representing the sum of the ranks of the absolute differences taking non-zero values obeys a normal distribution with mean $m = \frac{1}{2}(n_1(\frac{n_1+1}{2}) - S_0)$,

where S_0 is the sum of the ranks of the differences taking value 0, and variance given by one-fourth the sum of the squares of the ranks.

The Mann-Whitney and t -test described previously address the analysis of two groups of data; in case of three or more groups, the following tests may be used.

19.3.7 One-Way ANOVA

When comparing data across three or more groups, the obvious option of considering data one pair at a time presents itself. The problem with this approach is that it does not allow one to draw any conclusions about the dataset as a whole. While the probability that each individual pair yields significant results by mere chance is small, the probability that any one pair of the entire dataset does so is substantially larger. The One-Way ANOVA takes a comprehensive approach in analyzing data and attempts to extend the logic of t -tests to handle three or more groups concurrently. It uses the mean of the sum of squared deviates (SSD) as an aggregate measure of variability between and within groups. NOTE: For a sample of n observations X_1, X_2, \dots, X_n , the sum of squared deviates is given by

$$SSD = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}$$

The numerator in the t -statistic is representative of the difference in the mean **between** the two groups under scrutiny, while the denominator is a measure of the random variance **within** each group. For a dataset with k groups of size n_1, n_2, \dots, n_k , and mean values M_1, M_2, \dots, M_k respectively, One-Way ANOVA employs the SSD between groups, SSD_{bg} , as a measure of variability in group mean values, and the SSD within groups, SSD_{wg} as representative of the randomness of values within groups. Here,

$$SSD_{bg} \equiv \sum_{i=1}^k n_i (M_i - M)^2$$

and

$$SSD_{wg} \equiv \sum_{i=1}^k SSD_i$$

with M being the average value over the entire dataset and SSD_i the SSD within group i . (Of course it follows that sum $SSD_{bg} + SSD_{wg}$ is exactly the total variability of the entire data).

Again drawing a parallel to the t -test, computation of the variance is associated with the number of degrees of freedom (df) within the sample, which as seen earlier is $n - 1$ in the case of an n -sized sample. One might then reasonably suppose that SSD_{bg} has $df_{bg} = k - 1$ degrees of freedom and SSD_{wg} , $df_{wg} = \sum_{i=1}^k n_i - 1$. The mean of the squared deviates (MSD) in each case provides a measure of the variance between and within groups respectively and is given by $MSD_{bg} = \frac{SSD_{bg}}{df_{bg}}$ and $MSD_{wg} = \frac{SSD_{wg}}{df_{wg}}$.

If the null hypothesis is false, then one would expect the variability between groups to be substantial in comparison to that within groups. Thus MSD_{bg} may be thought of in some sense as $MSD_{hypothesis}$ and MSD_{wg} as MSD_{random} . This evaluation is formalized through computation of the

$$F - ratio = \frac{MSD_{bg}/df_{bg}}{MSD_{wg}/df_{wg}}$$

It can be shown that the F -ratio obeys the F -distribution with degrees of freedom df_{bg}, df_{wg} ; thus p -values may be easily assigned.

The One-Way ANOVA assumes independent and random samples drawn from a normally distributed source. Additionally, it also assumes that the groups have approximately equal variances, which can be practically enforced by requiring the ratio of the largest to the smallest group variance to fall below a factor of 1.5. These assumptions are especially important in case of unequal group-sizes. When group-sizes **are** equal, the test is amazingly robust, and holds well even when the underlying source distribution is not normal, as long as the samples are independent and random. In the unfortunate circumstance that the assumptions stated above do not hold and the group sizes are perversely unequal, we turn to the Welch ANOVA for unequal variance case or Kruskal-Wallis test when the normality assumption breaks down.

19.3.8 Post hoc testing of ANOVA results

The significant ANOVA result suggests rejecting the null hypothesis $H_0 =$ “means are the same”. It does not tell which means are significantly different. For a given gene, if any of the group pair is significantly different, then in ANOVA test the null hypothesis will be rejected. Post hoc tests are multiple comparison procedures commonly used on only those genes that are significant in ANOVA F-test. If the F-value for a factor turns out non significant, one cannot go further with the analysis. This ‘protects’ the post hoc test from being (ab)used too liberally. They are designed to keep the experiment wise error rate to acceptable levels.

The most common post hoc test is **Tukey’s** Honestly Significant Difference or HSD test . Tukey’s test calculates a new critical value that can be used to evaluate whether differences between any two pairs of means are significant. One simply calculates one critical value and then the difference between all possible pairs of means. Each difference is then compared to the Tukey critical value. If the difference is larger than the Tukey value, the comparison is significant. The formula for the critical value is:

$HSD = q\sqrt{\frac{MS_{error}}{n}}$, where q is the studentized range statistic (similar to the t-critical values, but different). MS_{error} is the mean square error from the overall F-test, and n is the sample size for each group. Error df is the df used in the ANOVA test.

SNK test is a less stringent test compared to Tukey HSD. $SNK = q_r\sqrt{\frac{MS_{error}}{n}}$ Different cells have different critical values. The r value is obtained by taking the difference in the number of steps between cells and q_r is obtained from standard table. In Tukey HSD the q value is identical to the lowest q from the Newman-Keuls.

19.3.9 Unequal variance (Welch) ANOVA

ANOVA assumes that the populations from which the data came all have the same variance, regardless of whether or not their means are equal. Heterogeneity in variance among different groups can be tested using Levine's test (not available in **GeneSpring GX**). If the user suspect that the variance may not be equal and the number of samples in each group is not same, then Welch ANOVA should be done.

In Welch ANOVA, each group is weighted by the ratio of the number of samples and the variance of that group. If the variance of a group equals zero, the weight of that group is replaced by a large number. When all groups have zero variance and equal mean, the null hypothesis is accepted, otherwise for unequal means the null hypothesis is rejected.

19.3.10 The Kruskal-Wallis Test

The Kruskal-Wallis (KW) test is the non-parametric alternative to the One-Way independent samples ANOVA, and is in fact often considered to be performing "ANOVA by rank". The preliminaries for the KW test follow the Mann-Whitney procedure almost verbatim. Data from the k groups to be analyzed are combined into a single set, sorted, ranked and then returned to the original group. All further analysis is performed on the returned ranks rather than the raw data. Now, departing from the Mann-Whitney algorithm, the KW test computes the **mean** (instead of simply the sum) of the ranks for each group, as well as over the entire dataset. As in One-Way ANOVA, the sum of squared deviates between groups, SSD_{bg} , is used as a metric for the degree to which group means differ. As before, the understanding is that the groups means will not differ substantially in case of the null hypothesis. For a dataset with k groups of sizes n_1, n_2, \dots, n_k each, $n = \sum_{i=1}^k n_i$ ranks will be accorded. Generally speaking, apportioning these n ranks amongst the k groups is simply a problem in combinatorics. Of course SSD_{bg} will assume a different value for each permutation/assignment of ranks. It can be shown that the mean value for SSD_{bg} over all permutations is $(k-1)\frac{n(n-1)}{12}$. Normalizing the observed SSD_{bg} with this mean value gives us the H -ratio, and a rigorous method for assessment of associated p-values: The distribution of the

$$H - ratio = \frac{SSD_{bg}}{\frac{n(n+1)}{12}}$$

may be neatly approximated by the chi-squared distribution with $k - 1$ degrees of freedom.

19.3.11 The Repeated Measures ANOVA

Two groups of data with inherent correlations may be analyzed via the paired t -Test and Mann-Whitney. For three or more groups, the Repeated Measures ANOVA (RMA) test is used. The RMA test is a close cousin of the basic, simple One-Way independent samples ANOVA, in that it treads the same path, using the sum of squared deviates as a measure of variability between and within groups. However, it also takes additional steps to effectively remove extraneous sources of variability, that originate in pre-existing individual differences. This manifests in a third sum of squared deviates that is computed for each individual set or row of observations. In a dataset with k groups, each of size n ,

$$SSD_{ind} = \sum_{i=1}^n k(A_i - M)^2$$

where M is the sample mean, averaged over the entire dataset and A_i is the mean of the k values taken by individual/row i . The computation of SSD_{ind} is similar to that of SSD_{bg} , except that values are averaged over individuals or rows rather than groups. The SSD_{ind} thus reflects the difference in mean per individual from the collective mean, and has $df_{ind} = n - 1$ degrees of freedom. This component is removed from the variability seen within groups, leaving behind fluctuations due to "true" random variance. The F -ratio, is still defined as $\frac{MSD_{hypothesis}}{MSD_{random}}$, but while $MSD_{hypothesis} = MSD_{bg} = \frac{SSD_{bg}}{df_{bg}}$ as in the garden-variety ANOVA.

$$MSD_{random} = \frac{SSD_{wg} - SSD_{ind}}{df_{wg} - df_{ind}}$$

Computation of p-values follows as before, from the F -distribution, with degrees of freedom $df_{bg}, df_{wg} - df_{ind}$.

19.3.12 The Repeated Measures Friedman Test

As has been mentioned before, ANOVA is a robust technique and may be used under fairly general conditions, provided that the groups being assessed are of the same size. The non-parametric Kruskal

Wallis test is used to analyse independent data when group-sizes are unequal. In case of correlated data however, group-sizes are necessarily equal. What then is the relevance of the Friedman test and when is it applicable? The Friedman test may be employed when the data is collection of ranks or ratings, or alternately, when it is measured on a non-linear scale.

To begin with, data is sorted and ranked **for each individual or row** unlike in the Mann Whitney and Kruskal Wallis tests, where the entire dataset is bundled, sorted and then ranked. The remaining steps for the most part, mirror those in the Kruskal Wallis procedure. The sum of squared deviates between groups is calculated and converted into a measure quite like the H measure; the difference however, lies in the details of this operation. The numerator continues to be SSD_{bg} , but the denominator changes to $\frac{k(k+1)}{12}$, reflecting ranks accorded to each individual or row.

19.3.13 The N-way ANOVA

The N-Way ANOVA is used to determine the effect due to N parameters concurrently. It assesses the individual influence of each parameter, as well as their net interactive effect.

GeneSpring GX uses type-III sum of square (SS) in N-way ANOVA [47, 45]. This is equivalent to the method of weighted squares of means or complete least square method of Overall and Spiegel [?]. The type-III ss is defined as follows :

Let A and B be the factors, each having several levels. The complete effects model for these two factors is

$$y_{ijk} = \mu + a_i + b_j + t_{ij} + e_{ijk},$$

where y_{ijk} is the k -th observation in ij -th treatment group, μ is the grand mean, $a_i(b_j)$ is additive combination and t_{ij} is the interaction term and e_{ijk} is the error term, which takes into account of the variation in y that cannot be accounted for by the other four terms on the right hand side of the equation. The difference in residual sum of square (RSS) of the models

$$y_{ijk} = \mu + a_i + t_{ij} + e_{ijk},$$

and

$y_{ijk} = \mu + a_i + b_j + t_{ij} + e_{ijk}$, is the SS corresponding to factor B. Similarly, for other factors we take the difference of RSS of the model excluding that factor and the full model.

GeneSpring GX ANOVA can handle both balanced and unbalanced design, though only full factorial design is allowed. For more than three factors, terms only up to 3-way interaction is calculated, due to computational complexity. Moreover, **GeneSpring GX** calculates maximum 1000 levels, i.e., if the total number of levels for 3-way interaction model is more than 1000 (main + doublet + triplet), then **GeneSpring GX** calculates only up to 2-way interactions. Still if the number of levels is more than 1000 **GeneSpring GX** calculates only the main effects.

Full factorial designs with no replicate excludes the highest level interaction (with previous constraints) to avoid over fitting.

Missing values are handled in **GeneSpring GX** ANOVA. If for a condition, if more than one sample has values, then ANOVA handles them. But, if all the samples have missing values, then those values (entities) are excluded for p-value computation and a separate list titled 'Excluded Entities' is output at the end.

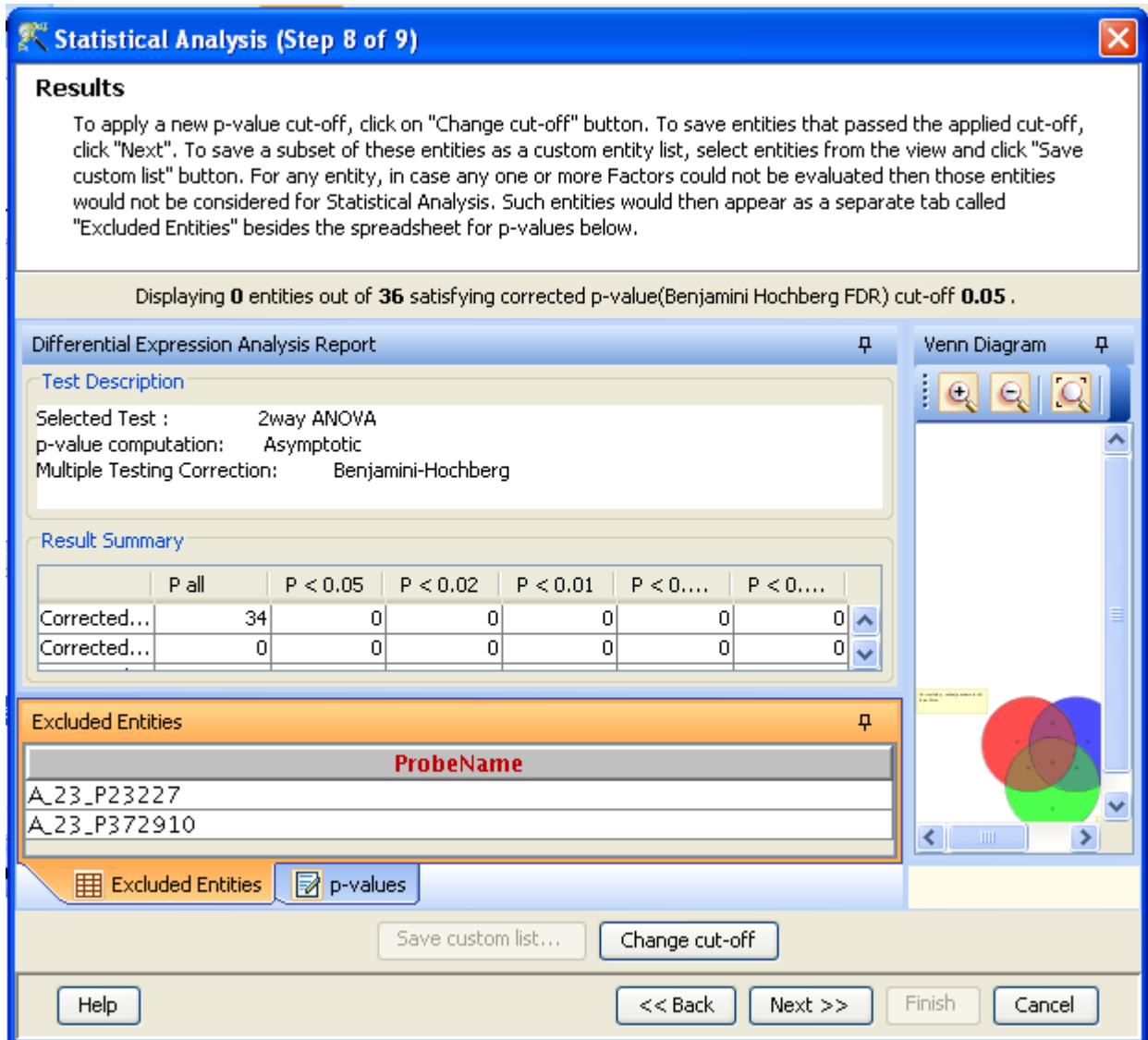


Figure 19.1: Anova result showing 'Excluded Entities' because of missing values

See Figure 19.1

19.4 Obtaining p -Values

Each statistical test above will generate a test value or statistic called the *test metric* for each gene. Typically, larger the test-metric more significant the differential expression for the gene in question. To identify all differentially expressed genes, one could just sort the genes by their respective test-metrics and then apply a cutoff. However, determining that cutoff value would be easier if the test-metric could be converted to a more intuitive p -value which gives the probability that the gene g appears as differentially

expressed purely by chance. So a p-value of .01 would mean that there is a 1% chance that the gene is not really differentially expressed but random effects have conspired to make it look so. Clearly, the actual p-value for a particular gene will depend on how expression values within each set of replicates are distributed. These distributions may not always be known.

Under the assumption that the expression values for a gene within each group are normally distributed and that the variances of the normal distributions associated with the two groups are the same, the above computed test-metrics for each gene can be converted into p -values, in most cases using closed form expressions. This way of deriving p -values is called *Asymptotic* analysis. However, if you do not want to make the normality assumptions, a *permutation analysis* method is sometimes used as described below.

19.4.1 p -values via Permutation Tests

As described in Dudoit et al. [21], this method does not assume that the test-metrics computed follows a certain fixed distribution.

Imagine a spreadsheet with genes along the rows and arrays along columns, with the first n_1 columns belonging to the first group of replicates and the remaining n_2 columns belonging to the second group of replicates. The left to right order of the columns is now shuffled several times. In each trial, the first n_1 columns are treated as if they comprise the first group and the remaining n_2 columns are treated as if they comprise the second group; the t -statistic is now computed for each gene with this new grouping. This procedure is ideally repeated $\binom{n_1+n_2}{n_1}$ times, once for each way of grouping the columns into two groups of size n_1 and n_2 , respectively. However, if this is too expensive computationally, a large enough number of random permutations are generated instead.

p -values for genes are now computed as follows. Recall that each gene has an actual test metric as computed a little earlier and several permutation test metrics computed above. For a particular gene, its p -value is the fraction of permutations in which the test metric computed is larger in absolute value than the actual test metric for that gene.

19.5 Adjusting for Multiple Comparisons

Microarrays usually have genes running into several thousands and tens of thousands. This leads to the following problem. Suppose p -values for each gene have been computed as above and all genes with a p -value of less than .01 are considered. Let k be the number of such genes. Each of these genes has a less than 1 in 100 chance of appearing to be differentially expressed by random chance. However, the chance that *at least* one of these k genes appears differentially expressed by chance is much higher than 1 in 100 (as an analogy, consider fair coin tosses, each toss produces heads with a 1/2 chance, but the chance of getting at least one heads in a hundred tosses is much higher). In fact, this probability could be as high $k * .01$ (or in fact $1 - (1 - .01)^k$ if the p -values for these genes are assumed to be independently distributed). Thus, a p -value of .01 for k genes does not translate to a 99 in 100 chance of all these genes

being truly differentially expressed; in fact, assuming so could lead to a large number of false positives. To be able to apply a p-value cut-off of .01 and claim that all the genes which pass this cut-off are indeed truly differentially expressed with a .99 probability, an adjustment needs to be made to these p -values.

See Dudoit et al. [21] and the book by Glantz [24] for detailed descriptions of various algorithms for adjusting the p -values. The simplest methods called the Holm step-down method and the Benjamini-Hochberg step-up methods are motivated by the description in the previous paragraph. **GeneSpring GX** offers 5 type of multiple correction, the first three corrects for Family-wise error rate (FWER) and the remaining ones correct False discovery rate (FDR). The fourth method, Benjamini-Yekutieli is only available in GO analysis.

1. Bonferroni correction
2. Bonferroni Step-down (Holm)
3. The Westfall-Young method
4. Benjamini-Yekutieli method
5. Benjamini-Hochberg method

The methods are listed in order of their stringency, with the Bonferroni being the most stringent, and the Benjamini and Hochberg FDR being the least stringent. The more stringent a multiple testing correction, the less false positive genes are allowed. The trade-off of a stringent multiple testing correction is that the rate of false negatives (genes that are called non-significant when they are) is very high.

In the examples, an error rate of 0.05 and a gene list of 1000 genes are assumed.

19.5.1 Bonferroni

Bonferroni method are single step procedure, where each p-value is corrected independently. The p-value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene will be significant.

Corrected P-value= p-value * n (number of genes in test) <0.05

As a consequence, if testing 1000 genes at a time, the highest accepted individual p-value is 0.00005, making the correction very stringent. With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.

19.5.2 Bonferroni Step-down (Holm method)

Holm's test is a stepwise method, also called a sequential rejection method, because it examines each hypothesis in an ordered sequence, and the decision to accept or reject the null hypothesis depends on the results of the previous hypothesis tests.

Genes are sorted in increasing order of p -value. The p -value of the j th gene in this order is now multiplied by $(n - j + 1)$ to get the new adjusted p -value. Because it is a little less corrective as the p -value increases, this correction is less conservative.

Example:

Gene Name	p-value before correction	Rank	Correction	Is gene significant after correction
A	0.00002	1	$0.00002 * 1000 = 0.02$	$0.02 < 0.05 \rightarrow Yes$
B	0.00004	2	$0.00004 * 999 = 0.039$	$0.039 < 0.05 \rightarrow Yes$
C	0.00009	3	$0.00009 * 998 = 0.0898$	$0.0898 > 0.05 \rightarrow No$

19.5.3 The Westfall-Young method

The Westfall and Young permutation method takes advantage of the dependence structure between genes, by permuting all the genes at the same time.

The Westfall and Young [51] procedure is a permutation procedure in which genes are first sorted by increasing t -statistic obtained on unpermuted data. Then, for each permutation, the test metrics obtained for the various genes in this permutation are artificially adjusted so that the following property holds: if gene i has a higher original test-metric than gene j , then gene i has a higher adjusted test metric for this permutation than gene j . The overall corrected p -value for a gene is now defined as the fraction of permutations in which the adjusted test metric for that permutation exceeds the test metric computed on the unpermuted data. Finally, an artificial adjustment is performed on the p -values so a gene with a higher unpermuted test metric has a lower p -value than a gene with a lower unpermuted test metric; this adjustment simply increases the p -value of the latter gene, if necessary, to make it equal to the former. Though not explicitly stated, a similar adjustment is usually performed with all other algorithms described here as well.

Because of the permutations, the method is very slow.

19.5.4 The Benjamini-Hochberg method

This method [7] assumes independence of p -values across genes. However, Benjamini and Yekuteili showed that the technical condition under which the test holds is that of positive regression dependency on each test statistics corresponding the true null hypothesis. In particular, the condition is satisfied by positively correlated normally distributed one sided test statistics and their studentized t -tests. Furthermore, since up-regulation and down-regulation are about equally likely to occur, the property of FDR control can be extended to two sided tests.

This procedure makes use of the ordered p -values $P_{(1)} \leq \dots \leq P_{(m)}$. Denote the corresponding null hypotheses $H_{(1)}, \dots, H_{(m)}$. For a desired FDR level q , the ordered p -value $P_{(i)}$ is compared to the critical value $q \cdot \frac{i}{m}$. Let $k = \max i : P_{(i)} \leq q \cdot \frac{i}{m}$. Then reject $H_{(1)}, \dots, H_{(k)}$, if such k exists.

19.5.5 The Benjamini-Yekutieli method

For more general cases, in which positive dependency conditions do not apply, Benjamini and Yekuteili showed that replacing q with $q / \sum_{i=1}^m (\frac{1}{i})$ will provide control of the FDR. This control is typically applied in GO analysis, since the GO terms have both positive and negative regression dependency.

19.5.6 Recommendations

1. The default multiple testing correction is the Benjamini and Hochberg False Discovery Rate. It is the least stringent of all corrections and provides a good balance between discovery of statistically significant genes and limitation of false positive occurrences.
2. The Bonferroni correction is the most stringent test of all, but offers the most conservative approach to control for false positives.
3. The Westfall and Young Permutation and Benjamini and Yekuteili are the only correction accounting for genes coregulation. However, Westfall and Young Permutation is slow and is also very conservative.

19.5.7 FAQ

1. **Q.** Why do I get more genes with a smaller gene list than with all genes list when I perform a one-way ANOVA using a Multiple Testing Correction?
A. As multiple testing corrections depend on how many genes are tested, the larger the gene list, the more stringent the correction will be. For instance, the Bonferroni correction will multiply the p-values of each gene by the number of tests performed. The more tests (or the more genes, since there is one test per gene), the smaller the p-value must be to pass the restriction.
2. **Q.** Why should I use a Multiple Testing Correction? If I select one, no genes pass the restriction.
A. Even though no genes pass the statistical restriction, it is important to keep in mind that genes that pass a restriction without multiple testing correction might all be false positives, thus not significant at all. If you have 10,000 genes in your genome, and perform a statistical analysis, a p-value cutoff of 0.05 allows a 5% chance of error. That means that 500 genes out of 10,000 could be found to be significant by chance alone.
3. **Q.** What should I do if no genes pass the statistical test when I apply the multiple testing correction?
A. To improve your statistical results, try one or more of the following suggestions:
 - Increase the p-value cutoff or error rate.
 - Increase the number of replicates in your experiment.
 - Select a smaller list of genes to use with your analysis. The smaller the list, the less stringent the multiple testing correction will be.

- Select a less stringent or no multiple testing correction. If you choose to apply no multiple testing correction, rank the genes by their p-values to inspect them manually. Genes with the smallest p-values will be the most reliable.
4. **Q.** When I increase the p-value cutoff, suddenly lot of genes passes a critical value. What is the reason for this ?
- A.** Typically this case arise when a permutative test is performed. If the the number of permutations are small then the minimum uncorrected p-value is large, say only 0.03. Hence a large number of genes can artificially have p-value 0.03 and when users increase p-value cutoff from 0.01 to 0.03 then those large number of genes will pass the cutoff.

Chapter 20

Clustering: Identifying Genes and Conditions with Similar Expression Profiles with Similar Behavior

20.1 What is Clustering

Cluster analysis is a powerful way to organize genes or entities and conditions in the dataset into clusters based on the similarity of their expression profiles. There are several ways of defining the [similarity measure](#), or the distance between two entities or conditions.

GeneSpring GX's clustering module offers the following unique features:

- A variety of clustering algorithms: [K-Means](#), [Hierarchical](#), and [Self Organizing Maps \(SOM\)](#), clustering, along with a variety of distance functions - [Euclidean](#), [Square Euclidean](#), [Manhattan](#), [Chebychev](#), [Differential](#), [Pearson Absolute](#), [Pearson Centered](#), and [Pearson Uncentered](#).

Data is sorted on the basis of such distance measures to group entities or conditions. Since different algorithms work well on different kinds of data, this large battery of algorithms and distance measures ensures that a wide variety of data can be clustered effectively.

- A variety of interactive views such as the [ClusterSet View](#), the [Dendrogram View](#), and the [U Matrix View](#) are provided for visualization of clustering results. These views allow drilling down into subsets of data and collecting together individual entity lists into new entity lists for further analysis. All views are lassoed, and enable visualization of a cluster in multiple forms based on the number of different views opened.
- The results of clustering algorithms are the following objects that are placed in the navigator and will be available in the experiment.

- Gene Tree: This is a [dendrogram](#) of the entities showing the relationship between the entities. This is a data object generated by Hierarchical Clustering.
- Condition Trees: This is a [dendrograms](#) of the conditions and shows the relationship between the conditions in the experiment. This is a data object generated by Hierarchical Clustering.
- Combined Trees: This is a two-dimensional [dendrograms](#) that results from performing *Hierarchical Clustering* on both entities and conditions which are grouped according to the similarity of their expression profiles.
- Classification: This is a [cluster set view](#) of entities grouped into clusters based on the similarity of their expression profiles.

20.2 Clustering Wizard

Running a clustering algorithm launches a wizard that allows users to specify the parameters required for the clustering algorithm and produces the results of clustering analysis. Upon examining the results of the chosen clustering algorithm you can choose to change the parameters and rerun the algorithm. If the clustering results are satisfactory, you can save the results as data objects in the analysis tree of the experiment navigator.

To perform Clustering analysis, click on the *Clustering* link within the *Analysis* section of the workflow panel.

Input parameters for clustering: In the first page of the clustering wizard, select the entity list, the interpretation and the clustering algorithm. By default, the active entity list and the active interpretation of the experiment is selected and shown in the dialog. To select a different entity list and interpretation for the analysis, click on the *Choose* button. This will show the tree of entity lists and interpretations in the current experiment. Select the entity list and interpretation that you would like to use for the analysis. Finally, select the clustering algorithm to run from the drop-down list and click *Next*. See Figure [20.1](#)

Clustering parameters In the second page of the clustering wizard, choose to perform clustering analysis on the selected entities, on conditions defined by the selected interpretations, or both entities and conditions. Select the distance measure from the drop-down menu. Finally, select the algorithm specific parameters. For details on the distance measures, refer the section of [distance measures](#). For details on individual clustering algorithms available in **GeneSpring GX**, see the following sections: [K-Means](#), [Hierarchical](#), [Self Organizing Maps \(SOM\)](#). Click *Next* to run the clustering algorithm with the selected parameters. See Figure [20.2](#)

Output views The third page of the clustering wizard shows the output views of the clustering algorithm. Depending on the parameters chosen and the algorithm chosen, the output views would be a combination of the following clustering views: [ClusterSet View](#), the [Dendrogram View](#), the and the [U Matrix View](#). These views allow users to visually inspect the quality of the clustering results. If the results are not satisfactory, click on the *Back* button, change the parameters and rerun the clustering algorithm. Once you are satisfied with the results, click *Next*. See Figure [20.3](#)

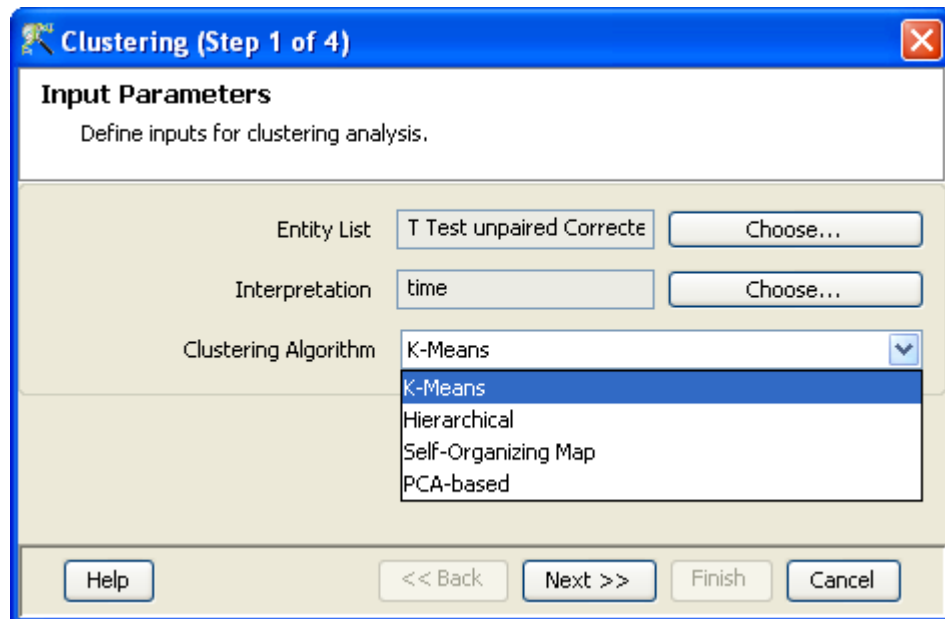


Figure 20.1: Clustering Wizard: Input parameters

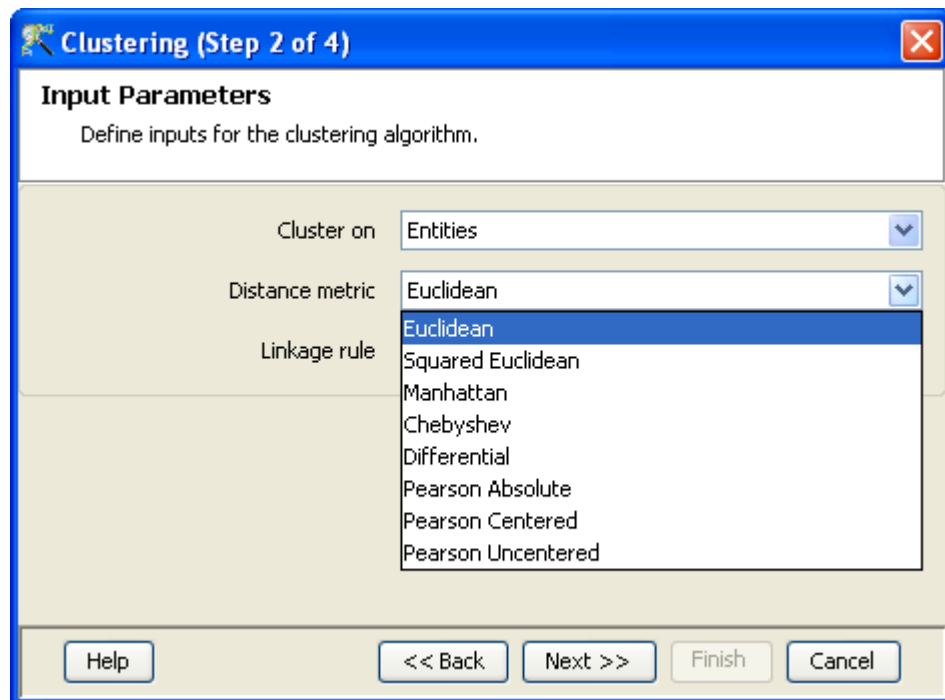


Figure 20.2: Clustering Wizard: Clustering parameters

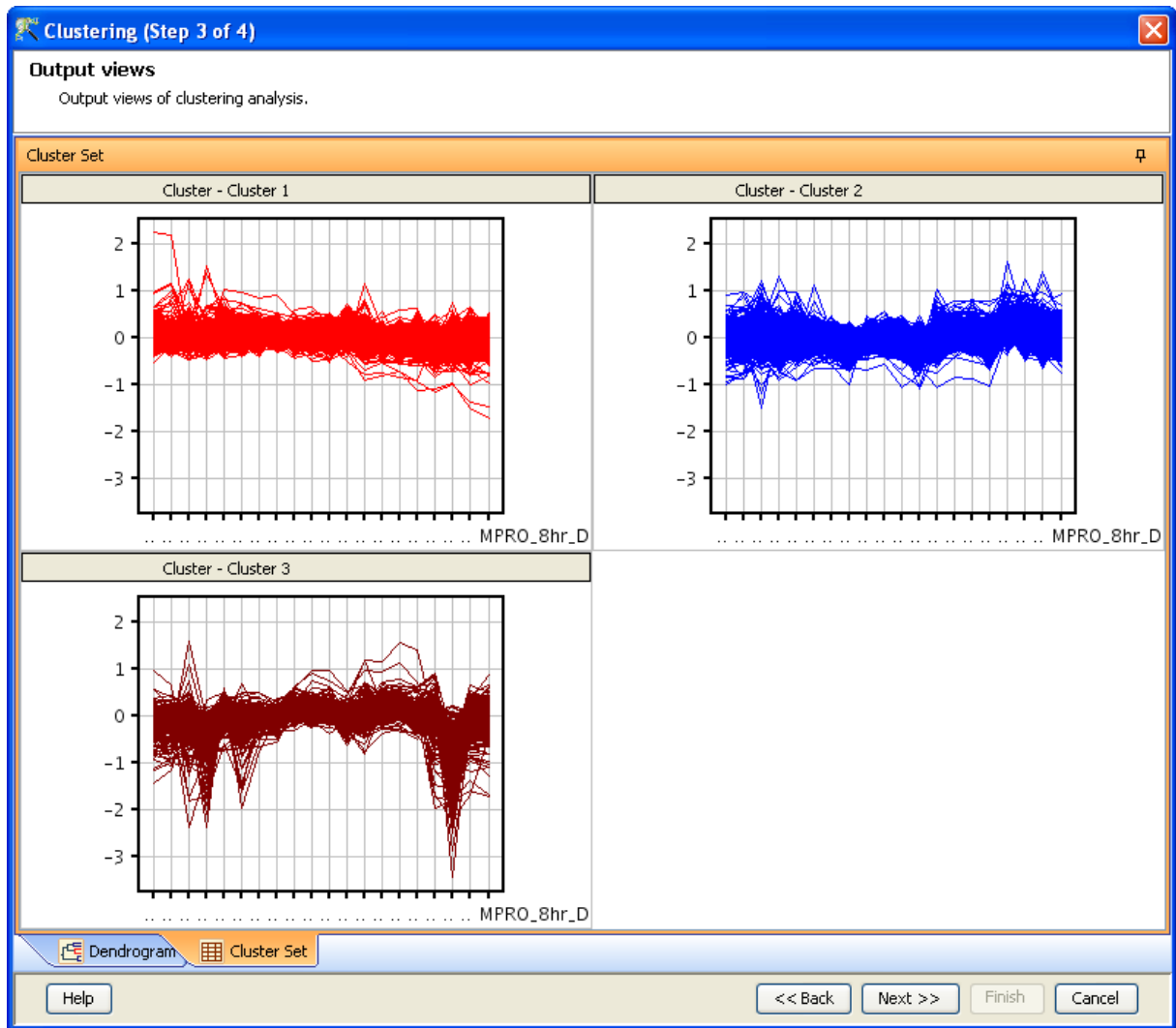


Figure 20.3: Clustering Wizard: Output Views

Object Details The final page of the clustering wizard shows the details of the result objects. It gives a default name to the object, and shows the parameters with which the clustering algorithm was run. You can change the name of the object and add notes to clustering object. Depending on the clustering algorithm, the objects would be a [classification object](#), [gene trees](#), [condition trees](#) or [combined trees](#). See [Figure 20.4](#)

20.3 Graphical Views of Clustering Analysis Output

GeneSpring GX incorporates a number of rich and intuitive graphical views of clustering results. All the views are interactive and allow the user to explore the results and create appropriate entity lists.

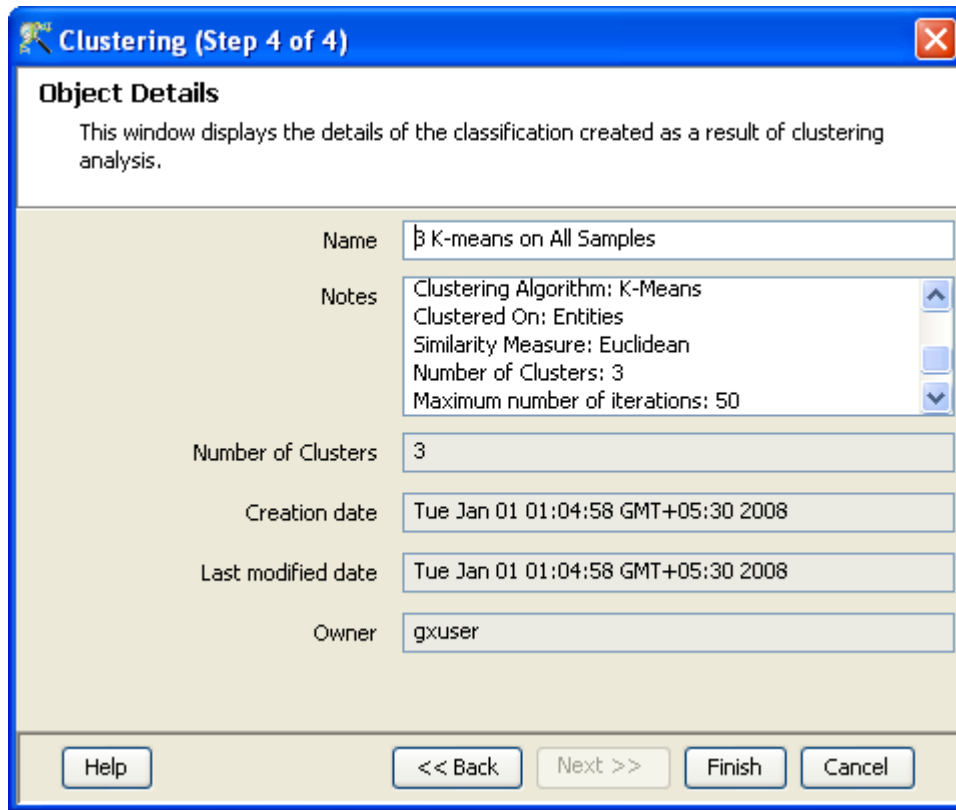


Figure 20.4: Clustering Wizard: Object details

20.3.1 Cluster Set or Classification

Algorithms like K-Means, SOM generate a fixed number of clusters. The Cluster Set plot graphically displays the profile of each clusters. Clusters are labelled as *Cluster 1*, *Cluster 2* ... and so on. See Figure 20.5

Cluster Set Operations

The Cluster Set view is a lassoed view and can be used to extract meaningful data for further use.

View Entities Profiles in a Cluster Double-click on an individual profile to bring up a entity inspector for the selected entity.

Create Entity Lists from Clusters: Once the classification object is saved in the *Analysis* tree, *Entity Lists* can be created from each cluster by right-clicking on the classification icon in the navigator and selecting *Expand as Entity List*.

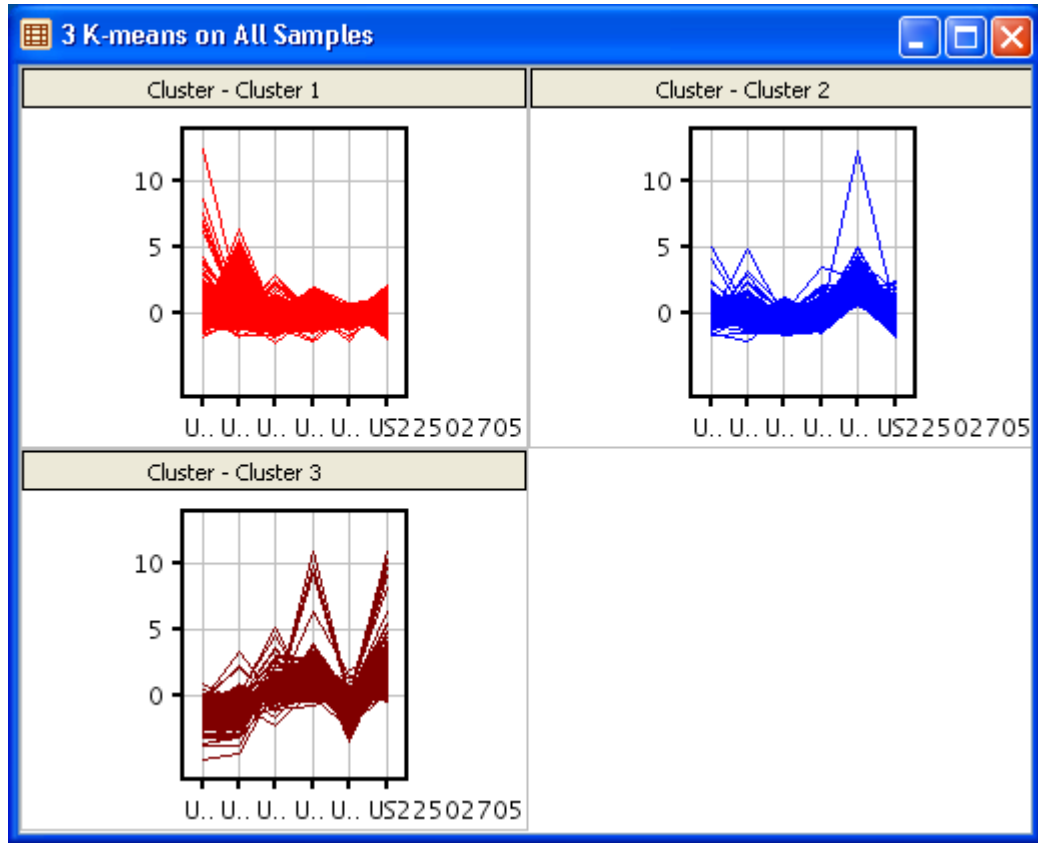


Figure 20.5: Cluster Set from K-Means Clustering Algorithm

Cluster Set Properties

The properties of the Cluster Set Display can be altered by right clicking on the *Cluster Set* view and choosing *Properties* from the drop-down menu.

The Cluster Set view, supports the following configurable properties:

Trellis The cluster set is a essentially *Profile Plot* trellised on the cluster. The number of rows and columns in the view can be changed from the *Trellis* tab of the dialog.

Axes The grids, axes labels, and the axis ticks of the plots can be configured and modified. To modify these, Right-Click on the view, and open the Properties dialog. Click on the Axis tab. This will open the axis dialog.

The plot can be drawn with or without the grid lines by clicking on the 'Show grids' option.

The ticks and axis labels are automatically computed and shown on the plot. You can show or remove the axis labels by clicking on the Show Axis Labels check box. Further, the orientation of the tick labels for the X-Axis can be changed from the default horizontal position to a slanted position or vertical position by using the drop down option and by moving the slider for the desired angle.

The number of ticks on the axis are automatically computed to show equal intervals between the minimum and maximum and displayed. You can increase the number of ticks displayed on the plot by moving the Axis Ticks slider. For continuous data columns, you can double the number of ticks shown by moving the slider to the maximum. For categorical columns, if the number of categories are less than ten, all the categories are shown and moving the slider does not increase the number of ticks.

Visualization Each cluster set can be assigned either a fixed customizable color or a color based on its value in a specified column. The *Customize* button can be used to customize colors.

In the cluster set plots, a mean profile can be drawn by selecting the box named *Display mean profile*.

Rendering The rendering of the fonts, colors and offsets on the *Cluster set* view can be customized and configured.

Fonts: All fonts on the plot can be formatted and configured. To change the font in the view, Right-Click on the view and open the Properties dialog. Click on the *Rendering* tab of the *Properties* dialog. To change a *Font*, click on the appropriate drop-down box and choose the required font. To customize the font, click on the customize button. This will pop-up a dialog where you can set the font size and choose the font type as bold or italic.

Special Colors: All the colors that occur in the plot can be modified and configured. The plot Background color, the Axis color, the Grid color, the Selection color, as well as plot specific colors can be set. To change the default colors in the view, Right-Click on the view and open the Properties dialog. Click on the Rendering tab of the Properties dialog. To change a color, click on the appropriate arrow. This will pop-up a *Color Chooser*. Select the desired color and click *OK*. This will change the corresponding color in the View.

Offsets: The bottom offset, top offset, left offset, and right offset of the plot can be modified and configured. These offsets may be need to be changed if the axis labels or axis titles are not completely visible in the plot, or if only the graph portion of the plot is required. To change the offsets, Right-Click on the view and open the Properties dialog. Click on the Rendering tab. To change plot offsets, move the corresponding slider, or enter an appropriate value in the text box provided. This will change the particular offset in the plot.

Quality Image The Profile Plot image quality can be increased by checking the High-Quality anti-aliasing option.

Columns The Profile Plot of each cluster is launched with the conditions in the interpretation. The set of visible conditions can be changed from the *Columns* tab. The columns for visualization and the order in which the columns are visualized can be chosen and configured for the column selector. Right-Click on the view and open the properties dialog. Click on the columns tab. This will open the column selector panel. The column selector panel shows the *Available items* on the left-side list box and the *Selected items* on the right-hand list box. The items in the right-hand list box are the columns that are displayed in the view in the exact order in which they appear.

To move columns from the *Available list* box to the *Selected list* box, highlight the required items in the *Available items* list box and click on the right arrow in between the list boxes. This will move the highlighted columns from the *Available items* list box to the bottom of the *Selected items* list box. To move columns from the Selected items to the *Available items*, highlight the required items on the *Selected items* list box and click on the left arrow. This will move the highlight columns from the *Selected items* list box to the *Available items* list box in the exact position or order in which the column appears in the experiment.

You can also change the column ordering on the view by highlighting items in the Selected items list box and clicking on the up or down arrows. If multiple items are highlighted, the first click will consolidate the highlighted items (bring all the highlighted items together) with the first item in the specified direction. Subsequent clicks on the up or down arrow will move the highlighted items as a block in the specified direction, one step at a time until it reaches its limit. If only one item or contiguous items are highlighted in the *Selected items* list box, then these will be moved in the specified direction, one step at a time until it reaches its limit. To reset the order of the columns in the order in which they appear in the experiment, click on the reset icon next to the *Selected items* list box. This will reset the columns in the view in the way the columns appear in the view.

To highlight items, Left-Click on the required item. To highlight multiple items in any of the list boxes, Left-Click and Shift-Left-Click will highlight all contiguous items, and Ctrl-Left-Click will add that item to the highlighted elements.

The lower portion of the Columns panel provides a utility to highlight items in the *Column Selector*. You can either match by *By Name* or *Column Mark* wherever appropriate. By default, the Match *By Name* is used.

- To match by Name, select Match By Name from the drop down list, enter a string in the Name text box and hit Enter. This will do a substring match with the *Available List* and the *Selected list* and highlight the matches.
- To match by Mark, choose Mark from the drop down list. The set of column marks (i.e., Affymetrix ProbeSet Id, raw signal, etc.) will be in the tool will be shown in the drop down list. Choose a Mark and the corresponding columns in the experiment will be selected.

Description The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

20.3.2 Dendrogram

Some clustering algorithms like Hierarchical Clustering do not distribute data into a fixed number of clusters, but produce a grouping hierarchy. Most similar entities are merged together to form a cluster and this combined entity is treated as a unit thereafter until all the entities are grouped together. The result is a tree structure or a dendrogram, where the leaves represent individual entities and the internal nodes represent clusters of similar entities.

The leaves are the smallest clusters with one entity or condition each. Each node in the tree defines a cluster. The distance at which two clusters merge (a measure of dissimilarity between clusters) is called the threshold distance, which is measured by the height of the node from the leaf. Every gene is labelled by its identifier as specified by the id column in the dataset.

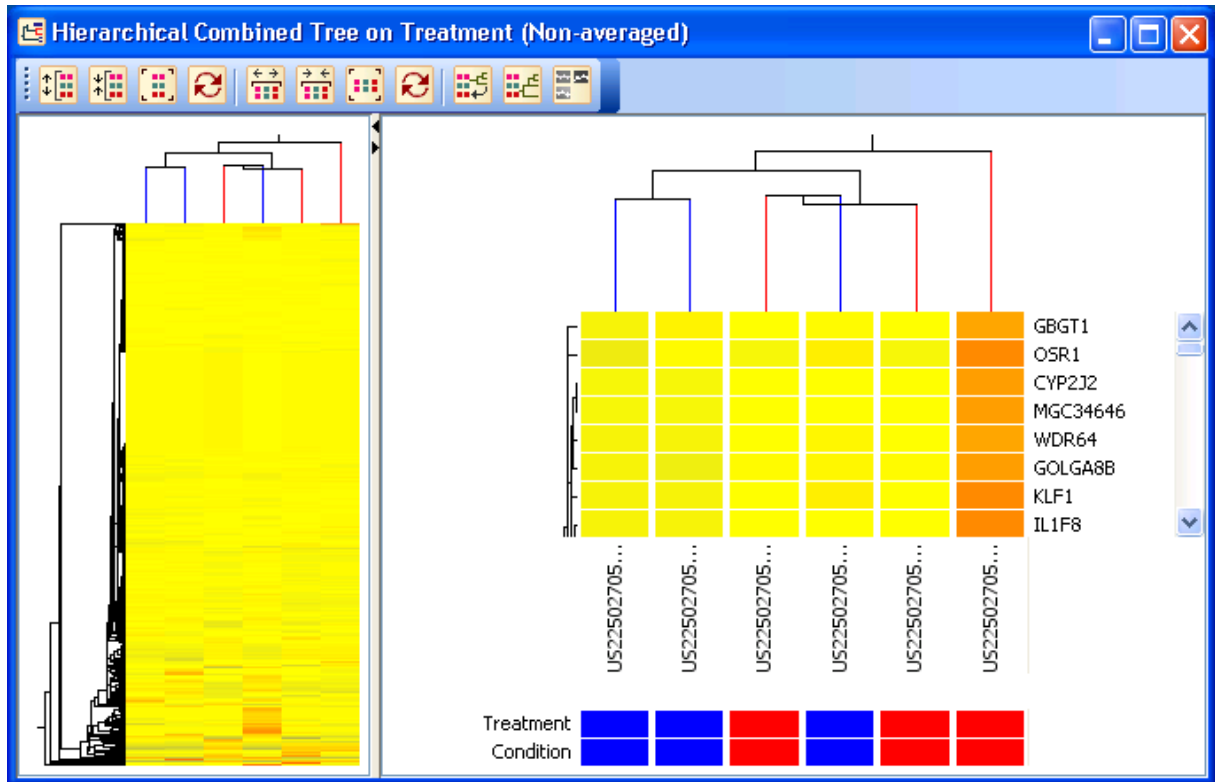


Figure 20.6: Dendrogram View of Clustering

The dendrogram view shows the tree in two panels. The left panel shows a bird's eye view of the whole tree and the right panel shows the expanded tree with scroll bars. If the number of rows are very large, the left panel intelligently samples the rows or columns and presents a bird's view of the whole dendrogram. Figure 20.6

The bottom of the left panel of the dendrogram, shows the condition color bar with the parameters in each interpretation.

When both entities and conditions are clustered, the plot includes two dendrograms - a vertical dendrogram for entities, and a horizontal one for conditions. This object is saved as a combined tree on the Analysis hierarchy in the navigator. The legend shows the color range of the heat map and the conditions on which clustering was performed.

When clustering is performed on entities, an entity tree object is created. When an entity tree view is launched, the tree is shown with all the entities on which the clustering was performed, with the columns of the active interpretation.

When clustering is performed on conditions in an experiment, a condition tree object is created. When a condition tree is launched, the tree is shown with the columns being the conditions on which clustering was performed, with the rows being the active entity list in the experiment.

Hovering over the cells of the heat map shows a tool-tip of the normalized intensity values. The tool-tip on the row header and the column header, shows the complete entity name or condition respectively. The tool-tip over the tree shows the distance values corresponding to the distance measure used in the clustering algorithm. The tool-tip on the condition color bar shows the conditions and the experimental parameter values for the interpretation.

Dendrogram Operations

The dendrogram is a lassoed view and can be navigated to get more detailed information about the clustering results. Dendrogram operations are also available by Right-Click on the canvas of the Dendrogram. Operations that are common to all views are detailed in the section [Common Operations on Table Views](#) above. In addition, some of the dendrogram specific operations are explained below:

Selection on Entity Trees and Condition Trees Left-Click on a cell in the heat map in either the panels selects the corresponding entity. Clicking on the row headers also selects entity.

Drawing a rectangle by left-click and dragging the mouse on the heat map in either of the panels, selects the entities (rows) and conditions (columns) corresponding to the cells that intersect the rectangle.

Click on the horizontal bar of entity to select the corresponding entity sub-tree. Click on the vertical bar of the condition tree to select the corresponding condition sub-tree.

The selected entities and conditions will be shown with the selection in both the panels of the dendrogram view and lassoed in all the view.

Click on the non-horizontal part of the entity tree to clear entity selection and click on the non-vertical portion of the condition tree to clear column selection.

Zoom Operations on Dendrogram The dendrogram can be zoomed into to view parts of condition trees and row trees. To zoom into a part of the dendrogram, draw a rectangle on the heat map by Shift-click and dragging the mouse on either panel of the dendrogram. The encompassing sub-tree containing the cells intersected by the drawn rectangle will be zoomed into and shown in the right panel. The tree node corresponding to the encompassing sub-tree will be shown with a blue dot. Thus zoomed portion could contain more cells than the cells intersected by the zoom window, since whole encompassing sub-tree will be shown in the right panel.

Shift-Click on the horizontal bar of entity to zoom into the corresponding entity sub-tree. Shift-Click on the vertical bar of the condition tree to zoom into the corresponding condition sub-tree.

Shift-Click on the non-horizontal part of the entity tree to reset zoom of the entity tree and shift-click on the non-vertical portion of the condition tree to reset zoom of of the condition tree.

Export As Image: This will pop-up a dialog to export the view as an image. This functionality allows the user to export very high quality image. You can choose to export only the visible region or export the whole image, by un-checking the *Export only visible region*. Exporting the whole image, will export the right panel of the dendrogram, showing the whole tree without the scroll bars.

You can specify any size of the image, as well as the resolution of the image by specifying the required dots per inch (dpi) for the image. Images can be exported in various formats. Currently supported

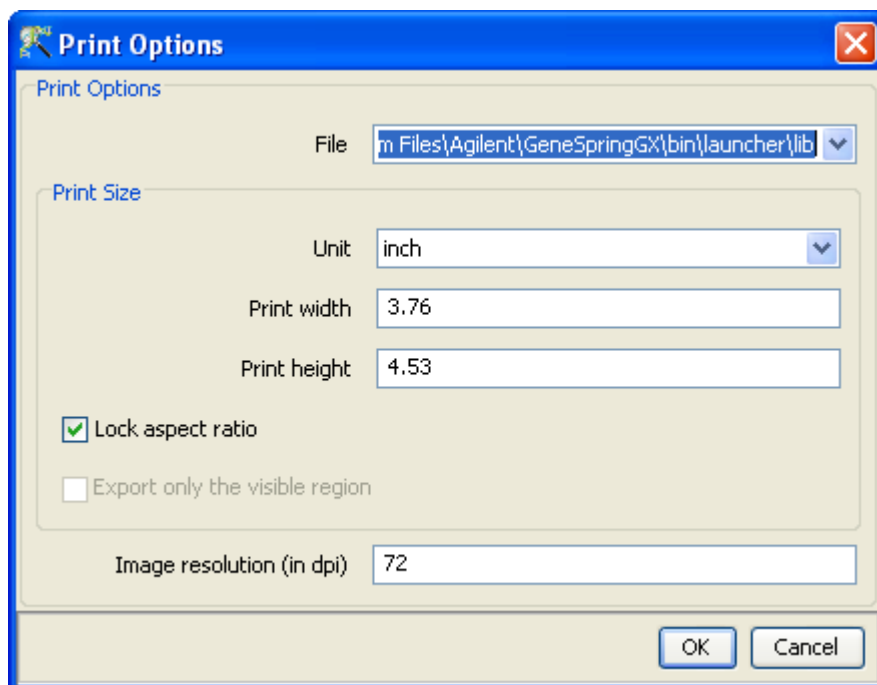


Figure 20.7: Export Image Dialog

formats include png, jpg, jpeg, bmp or tiff. Finally, images of very large size and resolution can be printed in the tiff format. Very large images will be broken down into tiles and recombined after all the images pieces are written out. This ensures that memory is not built up in writing large images. If the pieces cannot be recombined, the individual pieces are written out and reported to the user. However, tiff files of any size can be recombined and written out with compression. The default dots per inch is set to 300 dpi and the default size if individual pieces for large images is set to 4 MB. These default parameters can be changed in the *Tools* → *Options* → *Export as Image*. See Figure 20.7

Note: This functionality allows the user to create images of any size and with any resolution. This produces high-quality images and can be used for publications and posters. If you want to print vary large images or images of very high-quality the size of the image will become very large and will require huge resources. If enough resources are not available, an error and resolution dialog will pop us, saying the image is too large to be printed and suggesting you to try the tiff option, reduce the size of image or resolution of image, or to increase the memory available to the tool by changing the -Xmx option in `INSTALL_DIR/bin/packages/properties.txt` file. On **Mac OS X** the Java heap size parameters are set in in the file `Info.plist` located in `INSTALL_DIR/GeneSpringGX.app/Contents/Info.plist`. Change the Xmx parameter appropriately. Note that in the Java heap size limit on Mac OS X is about 2048M. See Figure 20.8

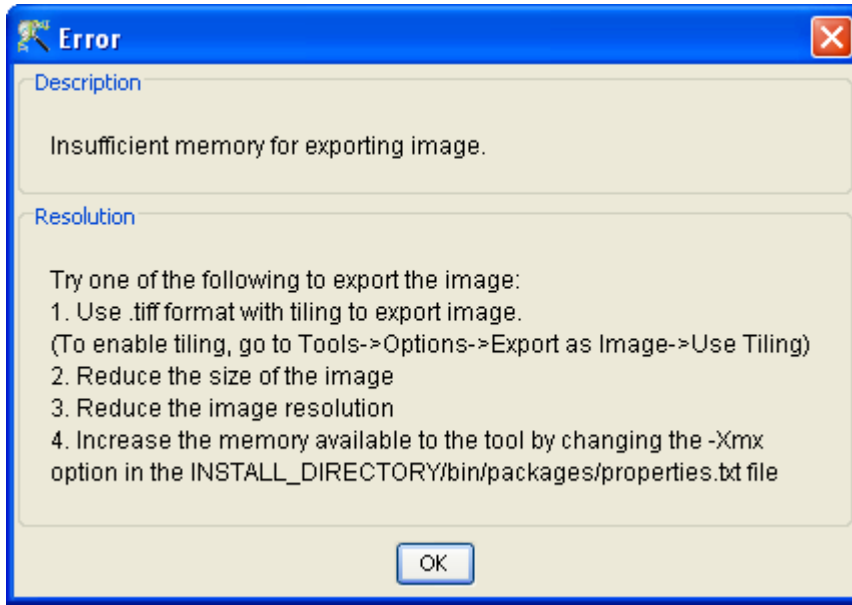


Figure 20.8: Error Dialog on Image Export



Figure 20.9: Dendrogram Toolbar

Note: You can export the whole dendrogram as a single image with any size and desired resolution. To export the whole image, choose this option in the dialog. The whole image of any size can be exported as a compressed tiff file. This image can be opened on any machine with enough resources for handling large image files.

Export as HTML: This will export the view as a html file. Specify the file name and the view will be exported as a HTML file that can be viewed in a browser and deployed on the web. If the whole image export is chosen, multiple images will be exported which is composed and opened in a browser.

Dendrogram Toolbar

The dendrogram toolbar offers the following functionality: See Figure 20.9



Expand rows: Click to increase the dimensions of the dendrogram. This increases the cell height in the right panel of the dendrogram. Row labels appear once the separation is large enough to accommodate label strings.



Contract rows: Click to reduce dimensions of the dendrogram. This decreases the cell height in the right panel of the dendrogram. Row labels appear only if the separation is large enough to accommodate label strings.



Fit rows to screen: This collapses the right panel of the dendrogram so that each cell is at least one pixel is size. If there are more rows that need to be accommodated, the right panel will be shown with a vertical scroll bar.



Reset rows: Click to scale the rows of the heat map back to default resolution.



Expand columns: Click to increase the dimensions of the dendrogram. This increases the cell width in the right panel of the dendrogram. Column labels appear once the separation is large enough to accommodate label strings.



Contract columns: Click to reduce dimensions of the dendrogram. This decreases the cell width in the right panel of the dendrogram. Column labels appear only if the separation is large enough to accommodate label strings.



Fit columns to screen: This collapses the right panel of the dendrogram so that each cell is at least one pixel is size. If there are more columns that need to be accommodated, the right panel will be shown with a horizontal scroll bar.



Reset rows: Click to scale the columns of the heat map back to default resolution.



Reset subtree: Click to reset the dendrogram in the right panel to show the whole tree



Save subtree: Click to save the subtree displayed in the right panel as a separate subtree. This will be saved in the navigation panel, in the *Analysis* folder under the appropriate entity list as a subtree object



Create classification: Clicking will launch a slider window with a ruler on the entity tree. Specify the threshold distance at which the classification object should be created. This will create a classification object with different entities in each cluster based upon the clustering results

Dendrogram Properties

The Dendrogram view supports the following configurable properties accessible from the right-click *Properties* dialog:

Visualization: Row headers: Any annotation column can be used to label the rows of the dendrogram from the **Row headers** drop down list.

Column headers: The column headers on the dendrogram is labeled with the names of the interpretation on which the heat map is launched. If all samples are used, or an unaveraged interpretation is used, the column headers show the column names. If column headers are not required, they can set to **None** from the drop-down list.

Color range: The Color and Saturation Threshold of the heat map can be changed from the Properties Dialog. The saturation threshold can be set by the Minimum, Center and Maximum sliders or by typing a numeric value into the text box and hitting Enter. The colors of Minimum, Center and Maximum can be set from the corresponding color chooser dialog. All values above the Maximum and values below the Minimum are thresholded to Maximum and Minimum colors respectively. The chosen colors are graded and assigned to cells based on the numeric value of the cell. Values between maximum and center are assigned a graded color in between the extreme maximum and center colors, and likewise for values between minimum and center.

Special Colors The color of the row tree and the tree highlight color of the dendrogram can be changed.

Rendering: The rendering of the dendrogram can be customized and configured from the rendering tab of the dendrogram properties dialog.

The location of the row and column headers can be set from the drop-down list.

The location of the row tree, the column tree and the condition bar can be changed from the drop-down list.

The row and column labels are shown along with the dendrogram. These widths allotted for these labels can be configured.

The width of the row tree and the height of the column tree can be changed.

The default vertical and horizontal spacing of the cells of the heat map can be changed.

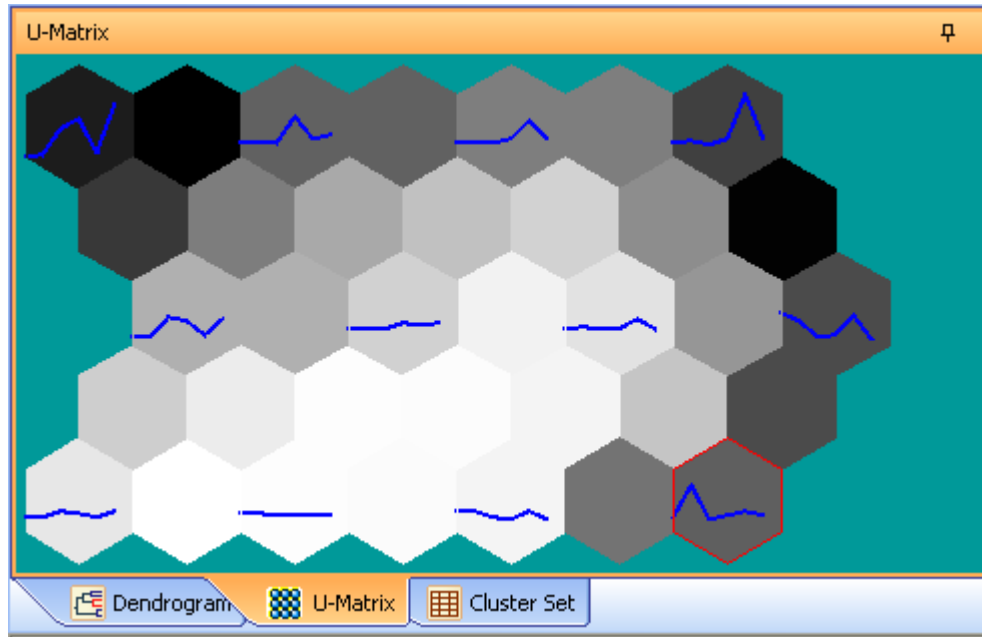


Figure 20.10: U Matrix for SOM Clustering Algorithm

Description: The title for the view and description or annotation for the view can be configured and modified from the description tab on the properties dialog. Right-Click on the view and open the Properties dialog. Click on the Description tab. This will show the Description dialog with the current Title and Description. The title entered here appears on the title bar of the particular view and the description if any will appear in the Legend window situated in the bottom of panel on the right. These can be changed by changing the text in the corresponding text boxes and clicking OK. By default, if the view is derived from running an algorithm, the description will contain the algorithm and the parameters used.

20.3.3 U Matrix

The U-Matrix view is used to display results of the SOM clustering algorithm. It is similar to the Cluster Set view, except that it displays clusters arranged in a 2D grid such that similar clusters are physically closer in the grid. The grid can be either hexagonal or rectangular as specified by the user. Cells in the grid are of two types, nodes and non-nodes. Nodes and non-nodes alternate in this grid. Holding the mouse over a node will cause that node to appear with a red outline. Clusters are associated only with nodes and each node displays the reference vector or the average expression profile of all entities mapped to the node. This average profile is plotted in blue. The purpose of non-nodes is to indicate the similarity between neighboring nodes on a grayscale. In other words, if a non-node between two nodes is very bright then it indicates that the two nodes are very similar and conversely, if the non-node is dark then the two nodes are very different. Further, the shade of a node reflects its similarity to its neighboring nodes. Thus not only does this view show average cluster profiles, it also shows how the various clusters are related. Left-clicking on a node will pull up the Profile plot for the associated cluster of entities. See Figure 20.10

U-Matrix Operations

The U-Matrix view supports the following operations.

Mouse Over Moving the mouse over a node representing a cluster (shown by the presence of the average expression profile) displays more information about the cluster in the tooltip as well as the status area. Similarly, moving the mouse over non-nodes displays the similarity between the two neighboring clusters expressed as a percentage value.

View Profiles in a Cluster Clicking on an individual cluster node brings up a *Profile Plot* view of the entities/conditions in the cluster. The entire range of functionality of the Profile view is then available.

U-Matrix Properties

The U-Matrix view supports the following properties which can be chosen by clicking *Visualization* under right-click *Properties* menu.

High quality image An option to choose high quality image. Click on *Visualization* under *Properties* to access this.

Description Click on *Description* to get the details of the parameters used in the algorithm.

20.4 Distance Measures

Every clustering algorithm needs to measure the similarity (difference) between entities or conditions. Once an entity or a condition is represented as a vector in n-dimensional expression space, several distance measures are available to compute similarity. **GeneSpring GX** supports the following distance measures:

- Euclidean: Standard sum of squared distance (L2-norm) between two entities.

$$\sqrt{\sum_i (x_i - y_i)^2}$$

- Squared Euclidean: Square of the Euclidean distance measure. This accentuates the distance between entities. Entities that are close are brought closer, and those that are dissimilar move further apart.

$$\sum_i (x_i - y_i)^2$$

- Manhattan: This is also known as the L1-norm. The sum of the absolute value of the differences in each dimension is used to measure the distance between entities.

$$\sum_i |x_i - y_i|$$

- Chebychev: This measure, also known as the L-Infinity-norm, uses the absolute value of the maximum difference in any dimension.

$$\max_i |x_i - y_i|$$

- Differential: The distance between two entities is estimated by calculating the difference in slopes between the expression profiles of two entities and computing the Euclidean norm of the resulting vector. This is a useful measure in time series analysis, where changes in the expression values over time are of interest, rather than absolute values at different times.

$$\sqrt{\sum_i [(x_{i+1} - x_i) - (y_{i+1} - y_i)]^2}$$

- Pearson Absolute: This measure is the absolute value of the Pearson Correlation Coefficient between two entities. Highly related entities give values of this measure close to 1, while unrelated entities give values close to 0.

$$\left| \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}} \right|$$

- Pearson Centered: This measure is the 1-centered variation of the Pearson Correlation Coefficient. Positively correlated entities give values of this measure close to 1; negatively correlated ones give values close to 0, and unrelated entities close to 0.5.

$$\frac{\left[\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}} \right) + 1 \right]}{2}$$

- Pearsons Uncentered This measure is similar to the Pearson Correlation coefficient except that the entities are not mean-centered. In effect, this measure treats the two entities as vectors and gives the cosine of the angle between the two vectors. Highly correlated entities give values close to 1, negatively correlated entities give values close to -1, while unrelated entities give values close to 0.

$$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

The choice of distance measure and output view is common to all clustering algorithms as well as other algorithms like *Find Similar Entities* algorithms in **GeneSpring GX**.

20.5 K-Means

This is one of the fastest and most efficient clustering techniques available, if there is some advance knowledge about the number of clusters in the data. Entities are partitioned into a fixed number (k) of clusters such that, entities/conditions within a cluster are similar, while those across clusters are dissimilar.

To begin with, entities/conditions are randomly assigned to k distinct clusters and the average expression vector is computed for each cluster. For every gene, the algorithm then computes the distance to all expression vectors, and moves the gene to that cluster whose expression vector is closest to it. The entire process is repeated iteratively until no entities/conditions can be reassigned to a different cluster, or a maximum number of iterations is reached. Parameters for K-means clustering are described below:

Cluster On Dropdown menu gives a choice of Entities, or Conditions, or Both entities and conditions, on which clustering analysis should be performed. Default is Entities.

Distance Metric Dropdown menu gives eight choices; Euclidean, Squared Euclidean, Manhattan, Chebyshev, Differential, Pearson Absolute, Pearson Centered, and Pearson Uncentered. The default is Euclidean.

Number of Clusters This is the value of k , and should be a positive integer. The default is 3.

Number of Iterations This is the upper bound on the maximum number of iterations for the algorithm. The default is 50 iterations.

Views The graphical views available with K-Means clustering are

- [Cluster Set View](#)

Advantages and Disadvantages of K-Means: K-means is by far the fastest clustering algorithm and consumes the least memory. Its memory efficiency comes from the fact that it does not need a distance matrix. However, it tends to cluster in circles, so clusters of oblong shapes may not be identified correctly. Further, it does not give relationship information for entities within a cluster or relationship information for the different clusters generated. When clustering with large datasets, use K-means to get smaller sized clusters and then run more computational intensive algorithms on these smaller clusters.

20.6 Hierarchical

Hierarchical clustering is one of the simplest and widely used clustering techniques for analysis of gene expression data. The method follows an agglomerative approach, where the most similar expression profiles are joined together to form a group. These are further joined in a tree structure, until all data forms a single group. The dendrogram is the most intuitive view of the results of this clustering method.

There are several important parameters, which control the order of merging entities and sub-clusters in the dendrogram. The most important of these is the linkage rule. After two most similar entities (clusters) are clubbed together, this group is treated as a single entity and its distances from the remaining groups (or entities) have to be re-calculated. **GeneSpring GX** gives an option of the following linkage rules on the basis of which two clusters are joined together:

Single Linkage: Distance between two clusters is the minimum distance between the members of the two clusters.

Complete Linkage: Distance between two clusters is the greatest distance between the members of the two clusters

Average Linkage: Distance between two clusters is the average of the pair-wise distance between entities in the two clusters.

Centroid Linkage: Distance between two clusters is the average distance between their respective centroids. This is the default linkage rule.

Ward's Method: This method is based on the ANOVA approach. It computes the sum of squared errors around the mean for each cluster. Then, two clusters are joined so as to minimize the increase in error.

Parameters for Hierarchical clustering are described below:

Cluster On Dropdown menu gives a choice of Entities, or Conditions, or Both entities and conditions, on which clustering analysis should be performed. Default is Entities.

Distance Metric Dropdown menu gives eight choices; Euclidean, Squared Euclidean, Manhattan, Chebyshev, Differential, Pearson Absolute, Pearson Centered, and Pearson Uncentered. The default is Euclidean.

Linkage Rule The dropdown menu gives the following choices; Complete, Single, Average, Centroid, and Wards. The default is Centroid linkage.

Views The graphical views available with Hierarchical clustering are

- [Dendrogram View](#)

Advantages and Disadvantages of Hierarchical Clustering: Hierarchical clustering builds a full relationship tree and thus gives a lot more relationship information than K-Means. However, it tends to connect together clusters in a local manner and therefore, small errors in cluster assignment in the early stages of the algorithm can be drastically amplified in the final result. Also, it does not output clusters directly; these have to be obtained manually from the tree.

20.7 Self Organizing Maps (SOM)

SOM Clustering is similar to K-means clustering in that it is based on a divisive approach where the input entities/conditions are partitioned into a fixed user defined number of clusters. Besides clusters, SOM produces additional information about the affinity or similarity between the clusters themselves by arranging them on a 2D rectangular or hexagonal grid. Similar clusters are neighbors in the grid, and dissimilar clusters are placed far apart in the grid.

The algorithm starts by assigning a random reference vector for each node in the grid. An entity/condition is assigned to a node, called the winning node, on this grid based on the similarity of its reference vector and the expression vector of the entity/condition. When an entity/condition is assigned to a node, the reference vector is adjusted to become more similar to the assigned entity/condition. The reference vectors of the neighboring nodes are also adjusted similarly, but to a lesser extent. This process is repeated iteratively to achieve convergence, where no entity/condition changes its winning node. Thus, entity/condition with similar expression vectors get assigned to partitions that are physically closer on the grid, thereby producing a topology that preserves the mapping from input space onto the grid.

In addition to producing a fixed number of clusters as specified by the grid dimensions, these proto-clusters (nodes in the grid) can be clustered further using hierarchical clustering, to produce a dendrogram based on the proximity of the reference vectors.

Cluster On Dropdown menu gives a choice of Entities, or Conditions, or Both entities and conditions, on which clustering analysis should be performed. Default is Entities.

Distance Metric Dropdown menu gives eight choices; Euclidean, Squared Euclidean, Manhattan, Chebyshev, Differential, Pearson Absolute, Pearson Centered, and Pearson Uncentered. The default is Euclidean.

Number of iterations This is the upper bound on the maximum number of iterations. The default value is 50.

Number of grid rows Specifies the number of rows in the grid. This value should be a positive integer. The default value is 3.

Number of grid columns Specifies the number of columns in the grid. This value should be a positive integer. The default value is 4.

Initial learning rate This defines the learning rate at the start of the iterations. It determines the extent of adjustment of the reference vectors. This decreases monotonically to zero with each iteration. The default value is 0.03.

Initial neighborhood radius This defines the neighborhood extent at the start of the iterations. This radius decreases monotonically to 1 with each iteration. The default value is 5.

Grid Topology This determines whether the 2D grid is hexagonal or rectangular. Choose from the dropdown list. Default topology is hexagonal.

Neighborhood type This determines the extent of the neighborhood. Only nodes lying in the neighborhood are updated when a gene is assigned to a winning node. The dropdown list gives two choices - Bubble or Gaussian. A Bubble neighborhood defines a fixed circular area, whereas a Gaussian neighborhood defines an infinite extent. However, the update adjustment decreases exponentially as a function of distance from the winning node. Default type is Bubble.

Run Batch SOM Batch SOM runs a faster simpler version of SOM when enabled. This is useful in getting quick results for an overview, and then normal SOM can be run with the same parameters for better results. Default is off.

Views The graphical views available with SOM clustering are

- [U-Matrix](#)
- [Cluster Set View](#)
- [Dendrogram View](#)

20.8 Missing Value Handling

For Clustering on rows, it requires each entity to have more than 50% of non-missing values across conditions. Entities having less than the required percentage of non-missing values are excluded from clustering. A separate entity list is created with valid entities and clustering is run on that list. In the case of Clustering on conditions it requires that each condition should have more than 50% of entities that have non-missing values. For Clustering on rows and conditions together, the above rules for rows and conditions are applied successively and in that order.

Chapter 21

Class Prediction: Learning and Predicting Outcomes

21.1 General Principles of Building a Prediction Model

Classification algorithms in **GeneSpring GX** are a set of powerful tools that allow researchers to exploit microarray data for building prediction models. These tools stretch the use of microarray technology into the arena of diagnostics and understanding the genetic basis of complex diseases. Classification predicts the class label of an input object. It requires an input data set, a subset of which is commonly known as training data, is used for creating a function for prediction of unknown class labels. A training data consists of input vector and an *answer* vector, and is used together with a learning method to train a knowledge database. The other subset is retained for subsequent use in confirming and validating the initial analysis. This set is commonly known as validation set.

Prediction models in **GeneSpring GX** build a model based on the expression profile of conditions. And with this model, try to predict the condition class of an unknown sample. For example, given gene expression data for different kinds of cancer samples, a model which can predict the cancer type for a new sample can be learnt from this data. **GeneSpring GX** provides a workflow link to build a model and predict the sample from gene expression data.

Model building for classification in **GeneSpring GX** is done using five powerful machine learning algorithms - [Decision Tree](#) (DT), [Neural Network](#) (NN), [Support Vector Machine](#) (SVM), [Naive Bayesian](#) (NB) and [PLSD](#) Models built with these algorithms can then be used to classify samples or genes into discrete classes based on its gene expression.

The models built by these algorithms range from visually intuitive (as with Decision Trees) to very abstract (as for Support Vector Machines). Together, these methods constitute a comprehensive toolset for learning, classification and prediction.

21.2 Prediction Pipeline

The problem statement for building a prediction model is to build a robust model to predict known phenotypic samples from gene expression data. This model is then used to predict an unknown sample based upon its gene expression characteristics. Here the model is built with the dependent variable being the sample type and the independent variable being the genes and their expression values corresponding to the sample. To cite the example stated above, given the gene expression profiles of the different types of cancerous tissue, you want to build a robust model, where, given the gene expression profile of a unknown sample, you will be able to predict the nature of the sample from the model. Thus the model must be generalizable and should work with a representative dataset. The model should not overfit the data used for building the model. In supervised learning

Once the model has been validated, the model can be saved and used to predict the outcome of a new sample from gene expression data of the sample. See Figure [21.1](#)

Note: All classification algorithms in **GeneSpring GX** for prediction of discrete classes (i.e. SVM, NN, NB,DT and PLSD) allow for validation, training and classification.

21.2.1 Validate

Validation helps to choose the right set of features or entity lists, an appropriate algorithm and associated parameters for a particular dataset. Validation is also an important tool to avoid over-fitting models on training data as over-fitting will give low accuracy on validation. Validation can be run on the same dataset using various algorithms and altering the parameters of each algorithm. The results of validation, presented in the [Confusion Matrix](#) (a matrix which gives the accuracy of prediction of each class), are examined to choose the best algorithm and parameters for the classification model.

Two types of validation have been implemented in **GeneSpring GX**.

Leave One Out: All data with the exception of one row is used to train the learning algorithm. The model thus learnt is used to classify the remaining row. The process is repeated for every row in the dataset and a Confusion Matrix is generated.

N-fold: The classes in the input data are randomly divided into N equal parts; N-1 parts are used for training, and the remaining one part is used for testing. The process repeats N times, with a different part being used for testing in every iteration. Thus each row is used at least once in training and once in testing, and a Confusion Matrix is generated. This whole process can then be repeated as many times as specified by the number of repeats.

The default values of three-fold validation and one repeat should suffice for most approximate analysis. If greater confidence in the classification model is desired, the Confusion Matrix of a 10-fold validation with

Classification Pipeline

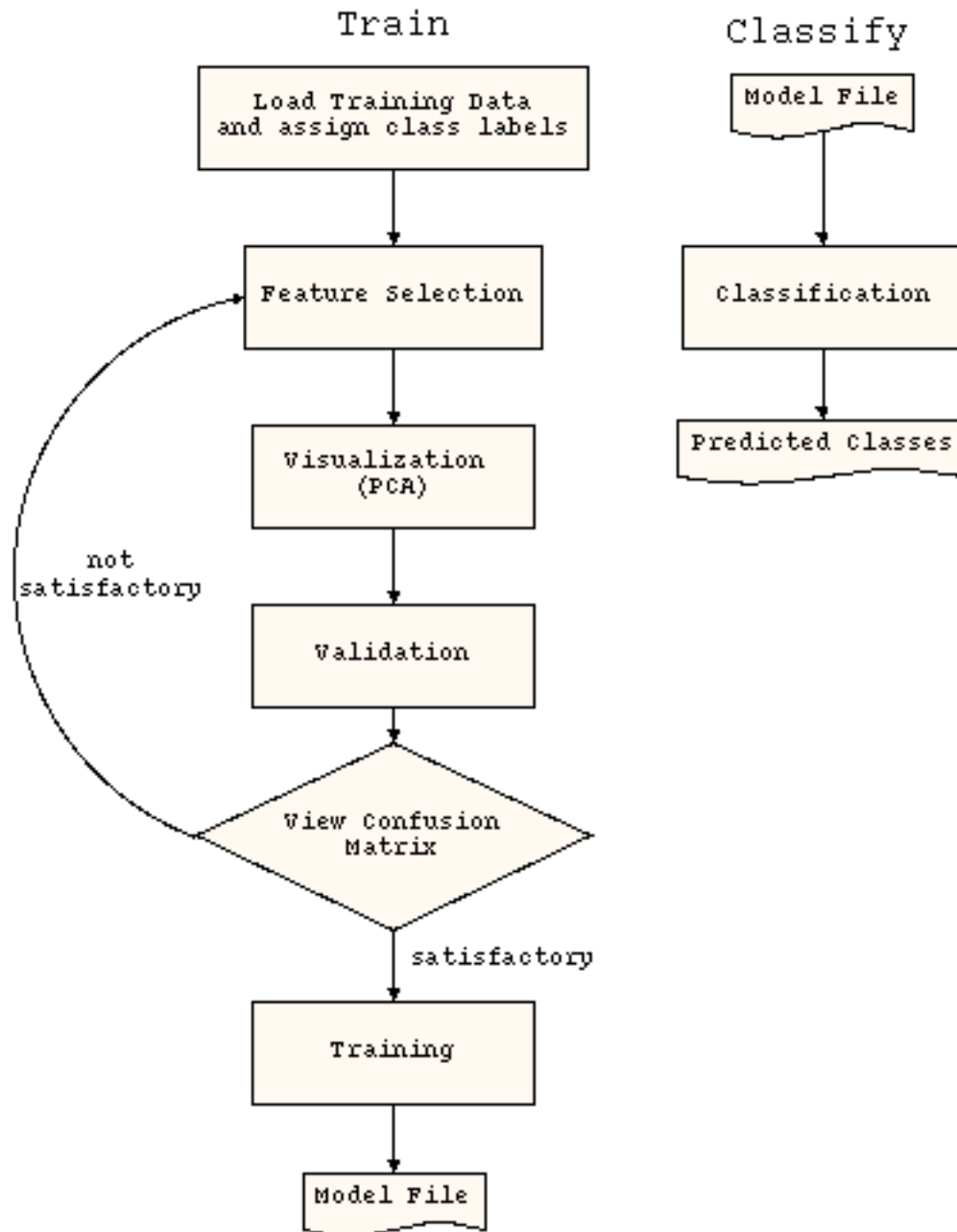


Figure 21.1: Classification Pipeline

three repeats needs to be examined. However, such trials would run the classification algorithm 30 times and may require considerable computing time with large datasets.

21.2.2 Prediction Model

Once the results of validation are satisfactory, as viewed from the confusion matrix of the validation process, a prediction model can be built and saved. The results of training yield a [Model](#), a [Report](#), a [Confusion Matrix](#) and a plot of the [Lorenz Curve](#). These views will be described in detail later.

21.3 Running Class Prediction in GeneSpring GX

Class prediction can be invoked from the workflow browser of the tool. There are two steps in class prediction; building prediction models and running prediction. Each of these takes you through a wizard collecting inputs providing visual outputs for examination and finally saving the results of building and running prediction models.

21.3.1 Build Prediction Model

The *Build Prediction Model* workflow link launches a wizard with five steps for building a prediction model.

Input Parameters The first step of building prediction models is to collect the required inputs. The prediction model is run on an entity list and an interpretation. The model is built to predict the interpretation based upon the expression values in the entity list. The entity list should thus be a filtered and analyzed entity list of genes that are significant to the interpretation. Normally these entity lists that are filtered and significant at a chosen p-value between the conditions in the interpretation. Thus the entity list is the set of features that are significant for the interpretation. See [Figure 21.2](#)

In the first step, the entity list, the interpretation and the class prediction algorithm are chose. By default, the entity list is the active entity list in the experiment. To change the entity list, click on the *Choose* button and select an entity list from the tree of entity list shown in the experiment. The default interpretation is the active interpretation in the dataset. To build a prediction model on another interpretation in the experiment, click on *Choose* and select another interpretation from the interpretation tree shown in the active experiment. Choose the prediction model from the drop-down list and click *Next*.

Validation Parameters The second step in building a prediction model is to choose the model parameters and the validation parameters. Here, the model specific parameters will be displayed and the validation type and parameters for validation can be chosen. For details on the model parameters see

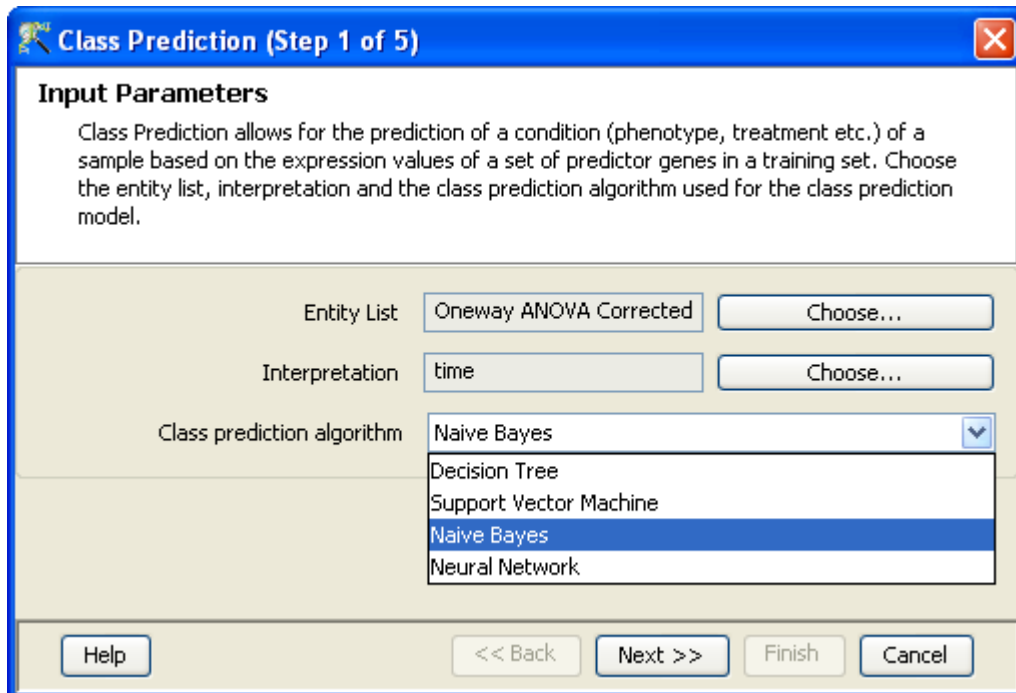


Figure 21.2: Build Prediction Model: Input parameters

the section on [Decision Tree \(DT\)](#), [Neural Network \(NN\)](#), [Support Vector Machine \(SVM\)](#), and [Naive Bayesian \(NB\)](#). For details on the validation parameters see the section on [Validate](#). See [Figure 21.3](#)

Validation Algorithm Outputs The next step in building prediction algorithms is to examine the validation algorithm outputs. These are a [confusion matrix](#) and a prediction [report table](#). The confusion matrix gives the efficacy of the prediction model and the report gives details of the prediction of each condition. For more details, see the section on [Viewing Classification Results](#). If the results are satisfactory, click *Next* or click *Back* to choose a different different model or a different set of parameters. Clicking *Next* will build the prediction model. See [Figure 21.4](#)

Training Algorithm Output The next step provides the output of the training algorithm. It provides a [confusion matrix](#) for the training model on the whole entity list, [report table](#), the [lorenz curve](#) showing the efficacy of classification and prediction model. Wherever appropriate, a visual output of the classification model is presented. For more details refer to the section on [Viewing Classification Results](#). For details on the model for each algorithm, go to the appropriate section. [Decision Tree \(DT\)](#), [Neural Network \(NN\)](#), [Support Vector Machine \(SVM\)](#), and [Naive Bayesian \(NB\)](#). If you want to rerun the model and change the parameters, click *Back*. Click *Next* to save the model. See [Figure 21.5](#)

Class Prediction Model Object The last step of building the prediction model is to save the class prediction model object in the tool. The view shows the model object with a default name and the notes showing the details of the prediction model and the parameters used. The view also shows a set of system generated fields that are stored with the model. You can change the name of the model and add additional notes in the text box provided. All these fields will be stored as annotations of the model can be searched and selected. Clicking *Finish* will save the model in the tool and show

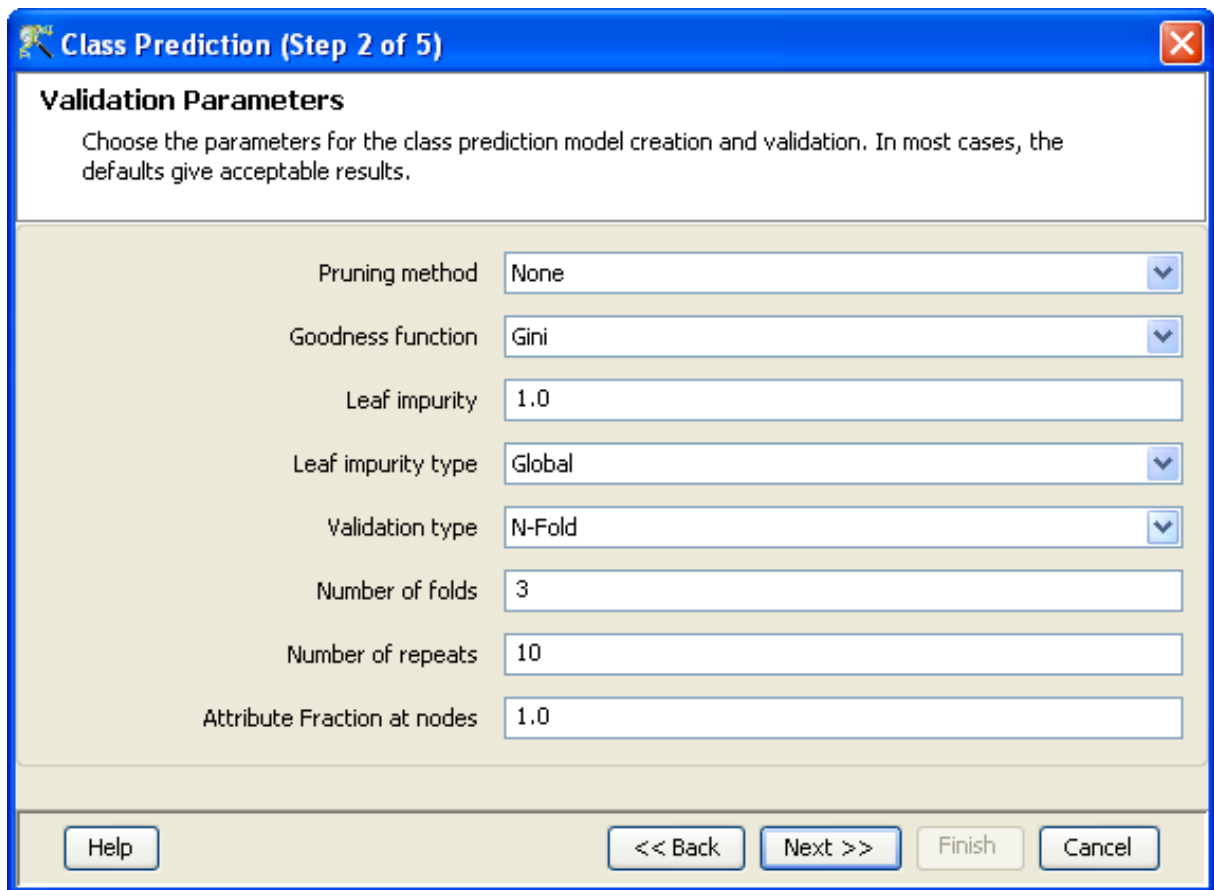


Figure 21.3: Build Prediction Model: Validation parameters

it in the *Analysis* tree of the experiment navigator. A right click on the model in the navigator will show options to inspect the model, copy it or remove it. Additionally, the entity list that was actually used in building the model can be created by clicking *Expand as Entity List*. This utility is useful to get that subset of the original entity list that actually goes into the model; this is especially true for decision trees where the final model is most likely to use a subset of the original entities.

The saved model can be used in any other experiment of the same technology in the tool. See Figure 21.6

21.3.2 Run Prediction

The *Run Prediction* workflow link is used to run a prediction model in an experiment. Clicking on this link will show all the models in the tool that have been created on the same technology. select a model and click *OK*. This will run the prediction model on the current experiment and output the results in a table. The model will take the entities in the technology used to model, run the model on all the samples in the experiment and predict the outcome for each sample in the experiment. The predicted results will be shown in the table along with a confidence measure appropriate to the model. For details on the prediction results and the confidence measures of prediction, see the appropriate sections [Decision Tree \(DT\)](#), [Neural](#)

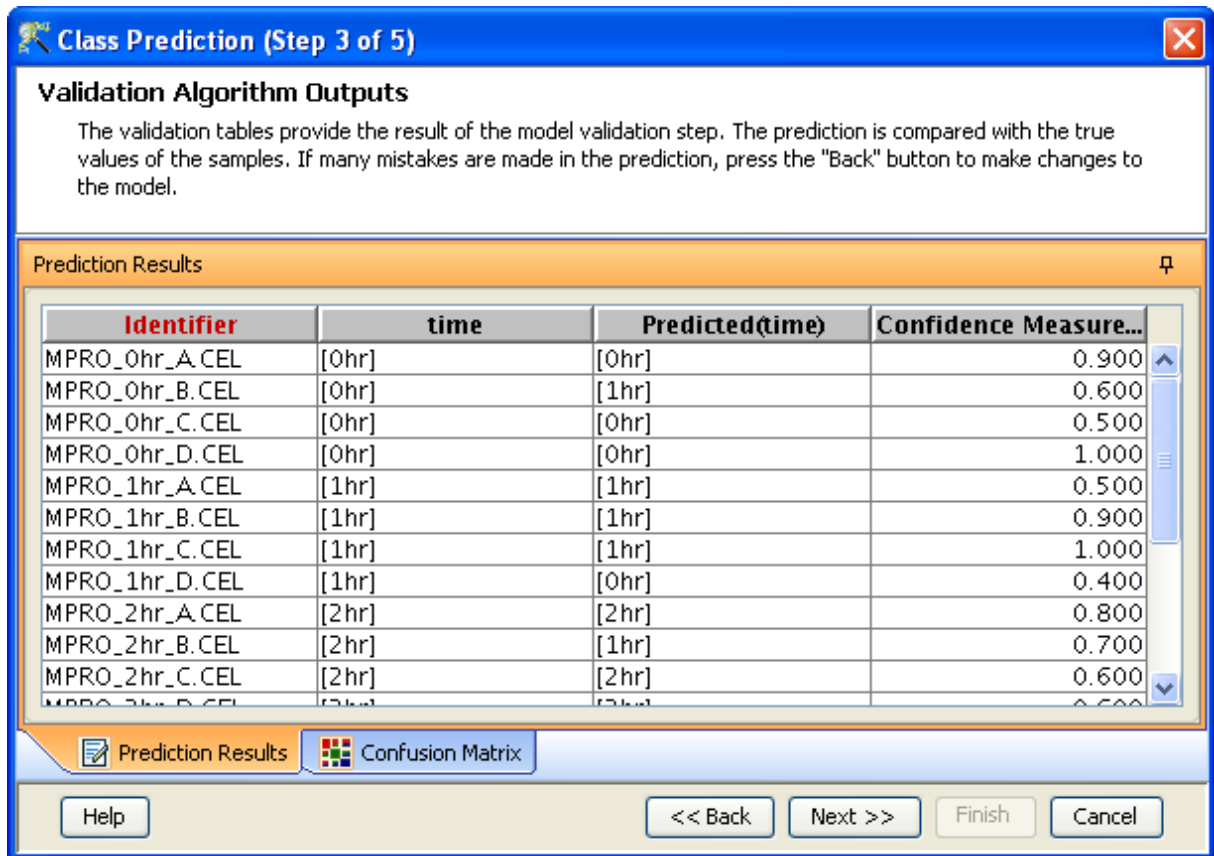


Figure 21.4: Build Prediction Model: Validation output

Network (NN), [Support Vector Machine \(SVM\)](#), and [Naive Bayesian \(NB\)](#). See [Figure 21.7](#)

Note: A prediction model created on a technology can be used only in experiments of the same technology.

21.4 Decision Trees

A Decision Tree is best illustrated by an example. Consider three samples belonging to classes A,B,C, respectively, which need to be classified, and suppose the rows corresponding to these samples have values shown below:

	Feature 1	Feature 2	Feature 3	Class Label
Sample 1	4	6	7	A
Sample 2	0	12	9	B
Sample 3	0	5	7	C

Table 21.1: Decision Tree Table

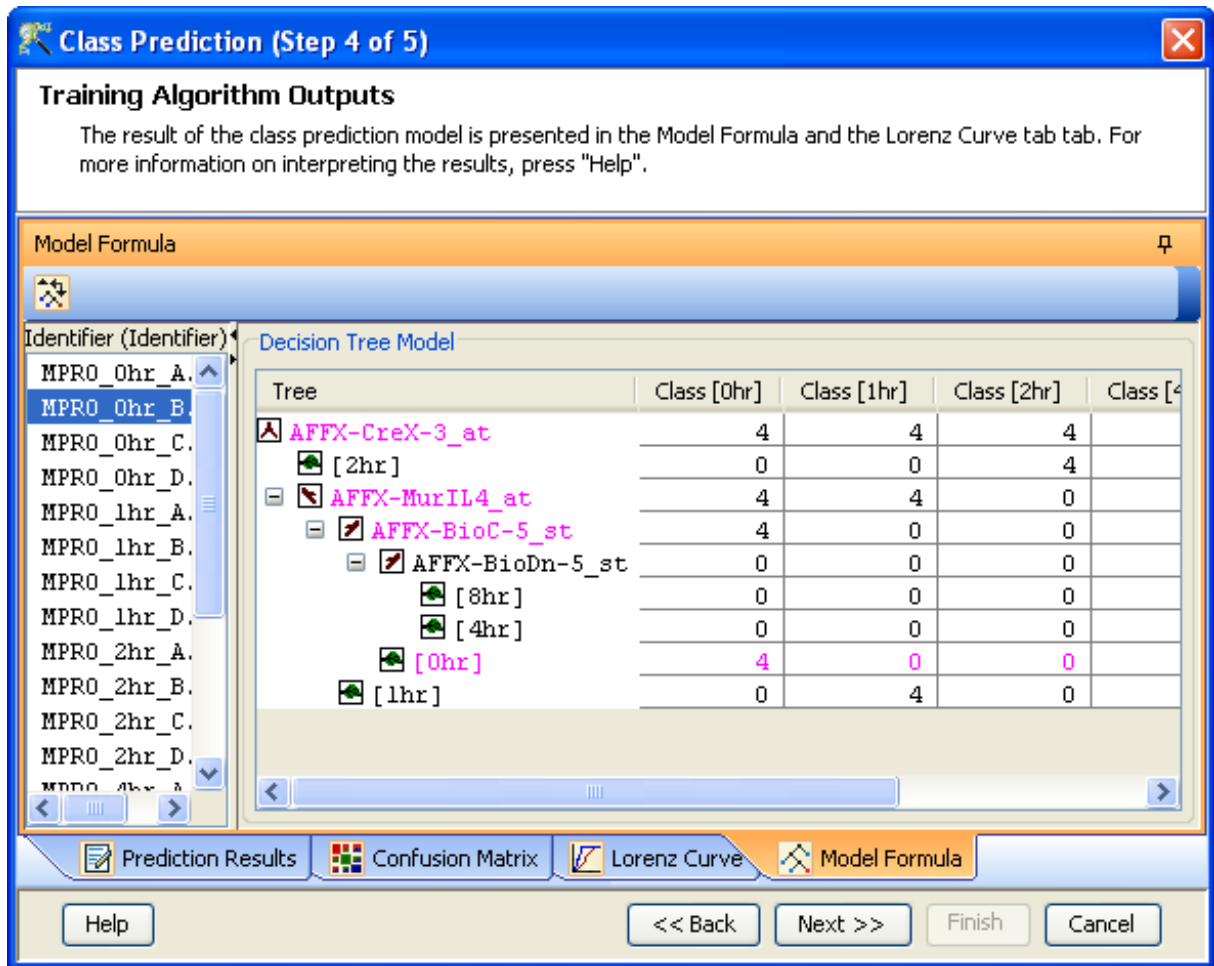


Figure 21.5: Build Prediction Model: Training output

Then the following sequence of Decisions classifies the samples - if feature 1 is at least 4 then the sample is of type A, and otherwise, if feature 2 is bigger than 10 then the sample is of Type B and if feature 2 is smaller than 10 then the sample is of type C. This sequence of if-then-otherwise decisions can be arranged as a tree. This tree is called a decision tree.

GeneSpring GX implements Axis Parallel Decision Trees. In an axis parallel tree, decisions at each step are made using one single feature of the many features present, e.g. a decision of the form if feature 2 is less than 10.

The decision points in a decision tree are called internal nodes. A sample gets classified by following the appropriate path down the decision tree. All samples which follow the same path down the tree are said to be at the same leaf. The tree building process continues until each leaf has purity above a certain specified threshold, i.e., of all samples which are associated with this leaf, at least a certain fraction comes from one class. Once the tree building process is done, a pruning process is used to prune off portions of the tree to reduce chances of over-fitting.

Class Prediction (Step 5 of 5)

Class Prediction Model

The information for the class prediction model is shown here. Press "Finish" to save the model.

Name	Naive Bayes model on celine
Notes	Created from Advanced Analysis operation: Build Prediction Model Experiment Name: MPRO Entity List: T Test unpaired Corrected p-valueP <= .05 Interpretation Name: celine
Creation date	Wed Dec 26 16:21:49 GMT+05:30 2007
Last modified date	Wed Dec 26 16:21:50 GMT+05:30 2007
Owner	gxuser
Technology	Affymetrix.GeneChip.MG_U74Av2
Algorithm Name	Naive Bayes
Overall Accuracy	0.75
Endpoint Name	celine
Number of Endpoints	4
Endpoint Value List	[[A], [B], [C], [D]]

Figure 21.6: Build Prediction Model: Model Object

The screenshot shows a software window titled "Output views of classification" with a sub-tab "Prediction Results". The window contains a table with the following data:

Identifier	celine	time	Predicted...	Confiden...
MPRO_0h...	A	0hr	[A]	1.000
MPRO_0h...	B	0hr	[A]	1.000
MPRO_0h...	C	0hr	[A]	1.000
MPRO_0h...	D	0hr	[D]	1.000
MPRO_1h...	A	1hr	[A]	1.000
MPRO_1h...	B	1hr	[B]	0.532
MPRO_1h...	C	1hr	[A]	1.000
MPRO_1h...	D	1hr	[D]	1.000
MPRO_2h...	A	2hr	[A]	1.000
MPRO_2h...	B	2hr	[B]	0.922
MPRO_2h...	C	2hr	[C]	0.868
MPRO_2h...	D	2hr	[D]	1.000
MPRO_4h...	A	4hr	[A]	1.000
MPRO_4h...	B	4hr	[B]	1.000
MPRO_4h...	C	4hr	[C]	1.000
MPRO_4h...	D	4hr	[D]	1.000
MPRO_8h...	A	8hr	[B]	0.812
MPRO_8h...	B	8hr	[B]	1.000
MPRO_8h...	C	8hr	[C]	0.815
MPRO_8h...	D	8hr	[B]	1.000

At the bottom of the window, there are tabs for "Model Formula" and "Prediction Results", and a "Close" button.

Figure 21.7: Run Prediction: Prediction output

Axis parallel decision trees can handle multiple class problems. Both varieties of decision trees produce intuitively appealing and visualizable classifiers.

21.4.1 Decision Tree Model Parameters

The parameters for building a *Decision Tree Model* are detailed below:

Pruning Method The options available in the dropdown menu are - Minimum Error, Pessimistic Error, and No Pruning. The default is Minimum Error. The No Pruning option will improve accuracy at the cost of potential over-fitting.

Goodness Function Two functions are available from the dropdown menu - Gini Function and Information Gain. This is implemented only for the Axis Parallel decision trees. The default is Gini Function.

Allowable Leaf Impurity Percentage (Global or Local) If this number is chosen to be x with the global option and the total number of rows is y , then tree building stops with each leaf having at most $x*y/100$ rows of a class different from the majority class for that leaf. And if this number is chosen to be x with the local option, then tree building stops with at most $x\%$ of the rows in each leaf having a class different from the majority class for that leaf. The default value is 1% and Global. Decreasing this number will improve accuracy at the cost of over-fitting.

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold.

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

The results of validation with Decision Trees are displayed in the dialog. They consist of the [Confusion Matrix](#) and the [Lorenz Curve](#). The Confusion Matrix displays the parameters used for validation. If the validation results are good these parameters can be used for training.

The results of model building with Decision Tree are displayed in the view. These consists of [Decision Tree model](#), a [Report](#), a [Confusion Matrix](#), and a [Lorenz Curve](#), all of which will be described later.

21.4.2 Decision Tree Model

GeneSpring GX implements the axis parallel decision trees.

The Decision Tree Model shows the learnt decision tree and the corresponding table. The left panel lists the row identifiers(if marked)/row indices of the dataset. The right panel shows the collapsed view of the tree. Clicking on the Expand/Collapse Tree icon in the toolbar can expand it. The leaf nodes are marked with the Class Label and the intermediate nodes in the Axis Parallel case show the Split Attribute.

To Expand the tree Click on an internal node (marked in brown) to expand the tree below it. The tree can be expanded until all the leaf nodes (marked in green) are visible. The table on the right gives information associated with each node.

The table shows the Split Value for the internal nodes. When a candidate for classification is propagated through the decision tree, its value for the particular split attribute decides its path. For values below the split attribute value, the feature goes to the left node, and for values above the split attribute, it moves to the right node. For the leaf nodes, the table shows the predicted Class Label. It also shows the distribution of features in each class at every node, in the last two columns. See [Figure 21.8](#)

To View Classification Click on an identifier to view the propagation of the feature through the decision tree and its predicted Class Label.

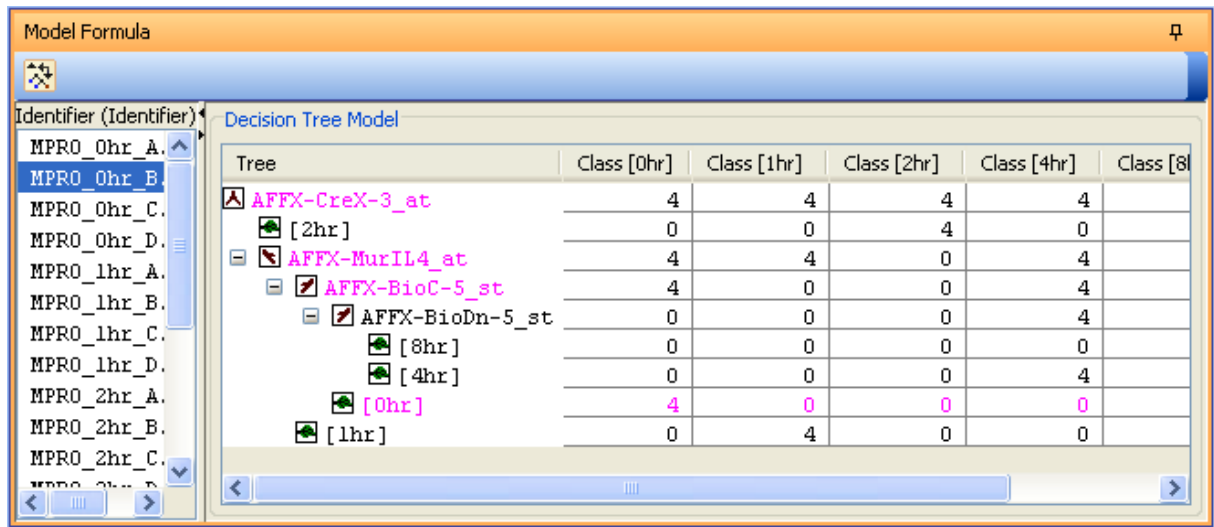


Figure 21.8: Axis Parallel Decision Tree Model



Expand/Collapse Tree: This is a toggle to expand or collapse the decision tree.

21.5 Neural Network

Neural Networks can handle multi-class problems, where there are more than two classes in the data. The Neural Network implementation in **GeneSpring GX** is the multi-layer perceptron trained using the back-propagation algorithm. It consists of layers of neurons. The first is called the input layer and features for a row to be classified are fed into this layer. The last is the output layer which has an output node for each class in the dataset. Each neuron in an intermediate layer is interconnected with all the neurons in the adjacent layers.

The strength of the interconnections between adjacent layers is given by a set of weights which are continuously modified during the training stage using an iterative process. The rate of modification is determined by a constant called the learning rate. The certainty of convergence improves as the learning rate becomes smaller. However, the time taken for convergence typically increases when this happens. The momentum rate determines the effect of weight modification due to the previous iteration on the weight modification in the current iteration. It can be used to help avoid local minima to some extent. However, very large momentum rates can also push the neural network away from convergence.

The performance of the neural network also depends to a large extent on the number of hidden layers (the layers in between the input and output layers) and the number of neurons in the hidden layers. Neural networks which use linear functions do not need any hidden layers. Nonlinear functions need at least one hidden layer. There is no clear rule to determine the number of hidden layers or the number of neurons

in each hidden layer. Having too many hidden layers may affect the rate of convergence adversely. Too many neurons in the hidden layer may lead to over-fitting, while with too few neurons the network may not learn.

21.5.1 Neural Network Model Parameters

The parameters for building a *Neural Network Model* are detailed below:

Number of Layers Specify the number of hidden layers, from layer 0 to layer 9. A value of '0' would mean 'no hidden layers'. In this case, the Neural Network behaves like a linear classifier. In **GeneSpring GX**, the default number of layers are 3.

Set Neurons This specifies the number of neurons in each layer. The default value is 15 neurons for each layer. Vary this parameter along with the number of layers.

Choose an optimal number of layers, which yield the best validation accuracy. Normally, up to 3 hidden layers are sufficient.

Number of Iterations The default is 100 iterations. This is normally adequate for convergence.

Learning Rate The default is a learning rate of 0.7. Decreasing this would improve chances of convergence but increase time for convergence.

Momentum The default is a 0.3.

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold validation in **GeneSpring GX**.

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

The results of validation with Neural Network are displayed in the dialog. They consist of the [Confusion Matrix](#) and the [Lorenz Curve](#). The Confusion Matrix displays the parameters used for validation. If the validation results are good these parameters can be used for training.

The results of training with Neural Network are displayed in the view. They consist of the [Neural Network model](#), a [Report](#), a [Confusion Matrix](#), and a [Lorenz Curve](#), all of which will be described later.

21.5.2 Neural Network Model

The Neural Network Model displays a graphical representation of the learnt model. There are two parts to the view. The left panel contains the row identifier(if marked)/row index list. The panel on the right contains a representation of the model neural network. The first layer, displayed on the left, is the input

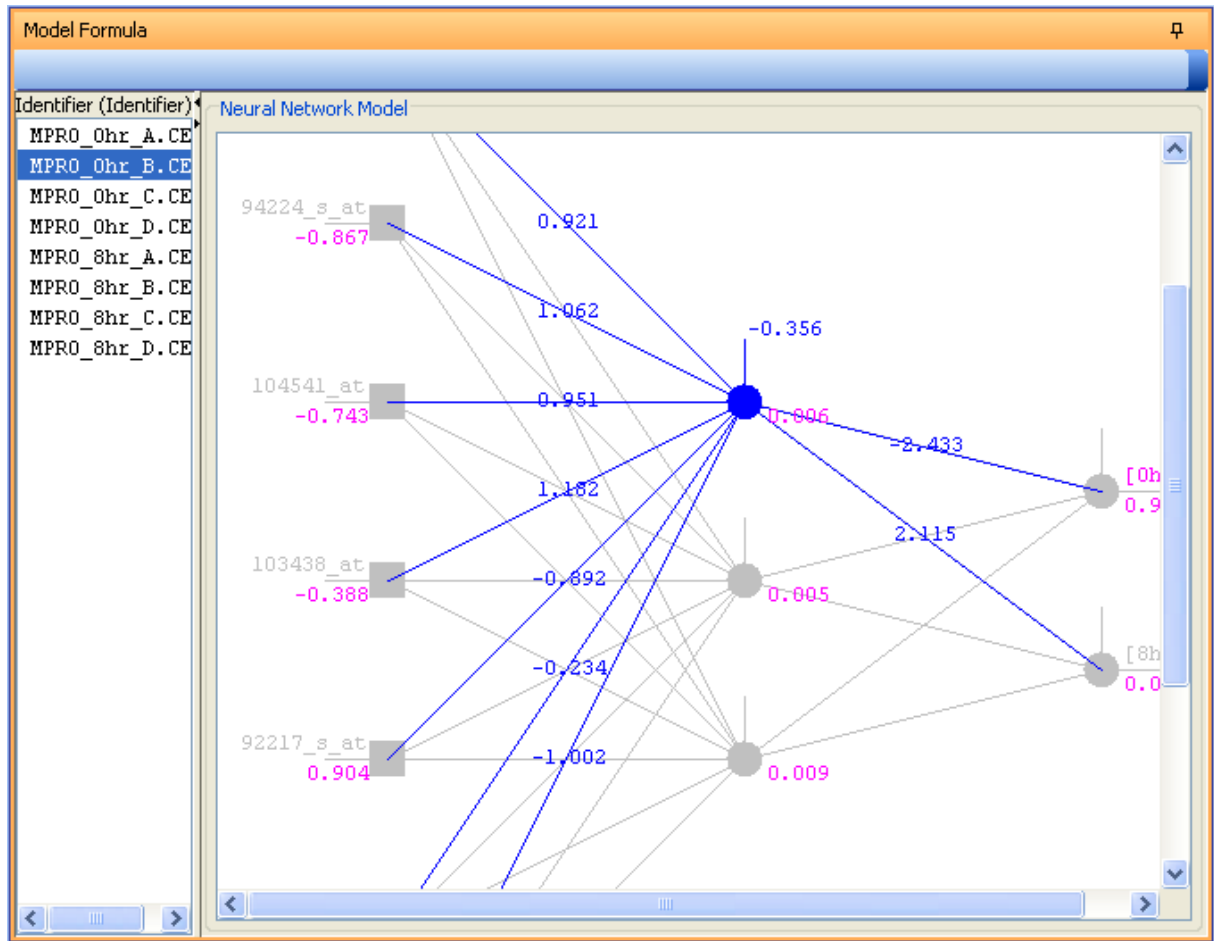


Figure 21.9: Neural Network Model

layer. It has one neuron for each feature in the dataset represented by a square. The last layer, displayed on the right, is the output layer. It has one neuron for each class in the dataset represented by a circle. The hidden layers are between the input and output layers, and the number of neurons in each hidden layer is user specified. Each layer is connected to every neuron in the previous layer by arcs. The values on the arcs are the weights for that particular linkage. Each neuron (other than those in the input layer) has a bias, represented by a vertical line into it. See Figure 21.9

To View Linkages Click on a particular neuron to highlight all its linkages in blue. The weight of each linkage is displayed on the respective linkage line. Click outside the diagram to remove highlights.

To View Classification Click on an id to view the propagation of the feature through the network and its predicted Class Label. The values adjacent to each neuron represent its activation value subjected to that particular input.

21.6 Support Vector Machines

Support Vector Machines (SVM) attempts to separate conditions or samples into classes by imagining these to be points in space and then determining a separating plane which separates the two classes of points.

While there could be several such separating planes, the algorithm finds a good separator which maximizes the separation between the classes of points. The power of SVMs stems from the fact that before this separating plane is determined, the points are transformed using a so called kernel function so that separation by planes post application of the kernel function actually corresponds to separation by more complicated surfaces on the original set of points. In other words, SVMs effectively separate point sets using non-linear functions and can therefore separate out intertwined sets of points.

The **GeneSpring GX** implementation of SVMs, uses a unique and fast algorithm for convergence based on the Sequential Minimal Optimization method. It supports three types of kernel transformations - Linear, Polynomial and Gaussian. In all these kernel functions, it so turns out that only the dot product (or inner product) of the rows (or conditions) is important and that the rows (or conditions) themselves do not matter, and therefore the description of the kernel function choices below is in terms of dot products of rows, where the dot product between rows a and b is denoted by $x(a).x(b)$.

The Linear Kernel is represented by the inner product given by the equation $x(a).x(b)$.

The Polynomial Kernel is represented by a function of the inner product given by the equation $(k_1[x(a).x(b)] + k_2)^p$, where p is a positive integer.

The Gaussian Kernel is given by the equation $e^{-\frac{(x(a)-x(b))^2}{\sigma}}$

Polynomial and Gaussian kernels can separate intertwined datasets but at the risk of over-fitting. Linear kernels cannot separate intertwined datasets but are less prone to over-fitting and therefore, more generalizable.

An SVM model consists of a set of support vectors and associated weights called Lagrange Multipliers, along with a description of the kernel function parameters. Support vectors are those points which lie on (actually, very close to) the separating plane itself. Since small perturbations in the separating plane could cause these points to switch sides, the number of support vectors is an indication of the robustness of the model; the more this number, the less robust the model. The separating plane itself is expressible by combining support vectors using weights called Lagrange Multipliers.

For points which are not support vectors, the distance from the separating plane is a measure of the belongingness of the point to its appropriate class. When training is performed to build a model, these belongingness numbers are also output. The higher the belongingness for a point, the more the confidence in its classification.

21.6.1 SVM ModelParameters

The parameters for building a *SVM Model* are detailed below:

Kernel Type Available options in the dropdown menu are - Linear, Polynomial, and Gaussian. The default is Linear.

Max Number of Iterations A multiplier to the number of conditions needs to be specified here. The default multiplier is 100. Increasing the number of iterations might improve convergence, but will take more time for computations. Typically, start with the default number of iterations and work upwards watching any changes in accuracy.

Cost This is the cost or penalty for misclassification. The default is 100. Increasing this parameter has the tendency to reduce the error in classification at the cost of generalization. More precisely, increasing this may lead to a completely different separating plane which has either more support vectors or less physical separation between classes but fewer misclassifications.

Ratio This is the ratio of the cost of misclassification for one class to the cost of the misclassification for the other class. The default ratio is 1.0. If this ratio is set to a value r , then the cost of misclassification for the class corresponding to the first row is set to the cost of misclassification specified, and the cost of misclassification for the other class is set to r times this value. Changing this ratio will penalize misclassification more for one class than the other. This is useful in situations where, for example, false positives can be tolerated while false negatives cannot. Then setting the ratio appropriately will have a tendency to control the number of false negatives at the expense of possibly increased false positives. This is also useful in situations where the classes have very different sizes. In such situations, it may be useful to penalize classifications much more for the smaller class than the bigger class.

Kernel Parameter (1) This is the first kernel parameter k_1 for polynomial kernels and can be specified only when the polynomial kernel is chosen. Default if 0.1.

Kernel parameter (2) This is the second kernel parameter k_2 for polynomial kernels. Default is set to 1. It is preferable to keep this parameter non-zero.

Exponent This is the exponent of the polynomial for a polynomial kernel (p). The default value is 2. A larger exponent increases the power of the separation plane to separate intertwined datasets at the expense of potential over-fitting.

Sigma This is a parameter for the Gaussian kernel. The default value is set to 1.0. Typically, there is an optimum value of sigma such that going below this value decreases both misclassification and generalization and going above this value increases misclassification. This optimum value of sigma should be close to the average nearest neighbor distance between points.

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold validation.

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

	Lagranges	Class Labels
0	0.257	NON-[B]
1	0.647	NON-[B]
2	1.954	NON-[B]
3	0.529	[B]
4	0.483	[B]
5	0.974	[B]
6	1.973	[B]
7	1.535	[B]
8	0.728	NON-[B]
9	1.333	NON-[B]
10	0.576	NON-[B]

Figure 21.10: Model Parameters for Support Vector Machines

The results of validation with SVM are displayed in the dialog. The Support Vector Machine view appears under the current spreadsheet and the results of validation are listed under it. They consist of the [Confusion Matrix](#) and the [Lorenz Curve](#). The Confusion Matrix displays the parameters used for validation. If the validations results are good then these parameters can be used for training.

The results of training with SVM are displayed in the dialog. They consist of the [SVM model](#), a [Report](#), a [Confusion Matrix](#), and a [Lorenz Curve](#), all of which will be described later.

Support Vector Machine Model

For [Support Vector Machine training](#), the model output contains the following training parameters in addition to the model parameters: See [Figure 21.10](#)

The top panel contains the Offset which is the distance of the separating hyperplane from the origin in addition to the input model parameters.

The lower panel contains the Support Vectors, with three columns corresponding to row identifiers(if marked)/row indices, Lagranges and Class Labels. These are input points, which determine the separating surface between two classes. For support vectors, the value of Lagrange Multipliers is non-zero and for other points it is zero. If there are too many support vectors, the SVM model has over-fit the data and may not be generalizable.

21.7 Naive Bayesian

Bayesian classifiers are parameter based statistical classifiers. They are multi-class classifiers and can handle continuous and categorical variables. They predict the probability that a sample belongs to a certain class. The Naive Bayesian classifier assumes that the effect of an attribute on a given class is independent of the value of other attributes. This assumption is called the *class conditional independence*. The Naive Bayesian model is built based on the probability distribution function of the training data along each feature. The model is then used to classify a data point based on the learnt probability density functions for each class.

Each row in the data is presented as an n dimensional feature vector, $X = (x_1, x_2, \dots, x_n)$. If there are m classes, C_1, C_2, \dots, C_m . Given an unknown data sample X the classifier predicts that X belongs to the class having the highest posterior probability, conditioned on X . The Naive Bayesian assigns X to class C_i if and only if

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \leq j \leq m, j \neq i$$

Applying bayesian rule, and given the assumption of *class conditional independence*, the probability can be computed as

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

The Probabilities $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ is estimated from the training samples and forms the Naive Bayesian Model.

21.7.1 Naive Bayesian Model Parameters

The parameters for building a *Naive Bayesian Model* are detailed below:

Validation Type Choose one of the two types from the dropdown menu - Leave One Out, N-Fold. The default is N fold validation.

Number of Folds If N-Fold is chosen, specify the number of folds. The default value is 3.

Number of Repeats The default value is 10.

The results of validation with Naive Bayesian are displayed in the dialog. They consist of the [Confusion Matrix](#), [Validation Report](#) and the [Lorenz Curve](#). The Confusion Matrix displays the parameters used for validation. If the validations results are good these parameters can be used to train and build a model.

	[A]	[B]	[C]	[D]
Class distribution	0.25	0.25	0.25	0.25
Posterior prob...	0	1	0	0
AFFX-BioDn-5_...				
Mean	0.033	-0.075	-0.069	-0.01
Standard devi...	0.2	0.09	0.162	0.202
AFFX-CreX-3_...				
Mean	-0.033	-0.021	-0.04	0.023
Standard devi...	0.228	0.175	0.162	0.165
92610_at (con...				
Mean	-0.001	0.098	0.061	-0.064
Standard devi...	0.281	0.357	0.256	0.288

Figure 21.11: Model Parameters for Naive Bayesian Model

The results of the model with are displayed in the dialog. They consist of the [NB Model Formula](#), a [Report](#), a [Confusion Matrix](#), and a [Lorenz Curve](#), all of which will be described later.

21.7.2 Naive Bayesian Model View

For [Naive Bayesian training](#), the model output contains the row identifier(if marked)/row index on the left panel and the Naive Bayesian Model parameters in the right panel. The Model parameters consist of the Class Distribution for each class in the training data and parameters for each feature or column. For continuous features the parameters are the mean and standard deviation for the particular class and for categorical variables these are the proportion of each category in the particular class. See [Figure 21.11](#)

To View Classification Clicking on a row identifier/index highlights the classified class of the sample. It shows the computed posterior probability for the selected sample. The row will be classified into that class which shows the largest posterior probability.

21.8 Partial Least Square Discrimination

PLSD is an extension of the PLSR (Partial Least Square Regression) - a PLS version of LDA (Linear Discriminant Analysis). PLSD is useful when you need to predict a set of variables, and identify them as functional classes from a large number of independent variables (i.e., predictors).

Notions and Notations

The PLS model is developed from a training set of N observations (objects, cases, compounds, etc.) with K \mathbf{X} -variables denoted by $x_k (k = 1, \dots, K)$, and M \mathbf{Y} -classes $y_m (m = 1, 2, \dots, M)$. These training data form the two matrices \mathbf{X} and \mathbf{Y} of dimensions $(N \times K)$ and $(N \times M)$, respectively.

Later, predictions for new observations are made based on their \mathbf{X} -data. This gives predicted t-scores, loadings, and prediction results with confidence intervals.

21.8.1 PLS Model and Parameters

The goal of PLS regression is to predict \mathbf{Y} from \mathbf{X} and to describe their common structure.

PLS regression decomposes both \mathbf{X} and \mathbf{Y} as a product of a common set of orthogonal factors and specific loadings. So, the independent variables are decomposed as $\mathbf{X} = \mathbf{T}\mathbf{P}^T$ with $\mathbf{T}^T\mathbf{T} = \mathbf{I}$; where \mathbf{I} and \mathbf{P} are the identity and loading matrices. Likewise, \mathbf{Y} is estimated as $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T$; where \mathbf{B} is a diagonal matrix with "regression weights" as the diagonal elements and \mathbf{C} is the "weight matrix" of the dependent variables. The columns of \mathbf{T} are the latent vectors.

The dependent variables are predicted using the multivariate regression formula $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T = \mathbf{X}\mathbf{B}_{\text{PLS}}$; where $\mathbf{B}_{\text{PLS}} = (\mathbf{P}^{T+})\mathbf{B}\mathbf{C}^T$ and \mathbf{P}^{T+} is the Moore-Penrose pseudoinverse of \mathbf{P}^T .

Step 1 of 5: Input Parameters Select the entity list and the interpretation along with the algorithm (PLSD).

Step 2 of 5: Validation Parameters Select the Model Parameters from the Validation Parameters dialog, and then click **Next**. Refer to Table 21.2 for details.

Step 3 of 5: Validation Algorithm Outputs The results of validation with PLS are displayed in the dialog. They consist of the [Confusion Matrix](#) and the [Prediction Results](#). If the validation results are good, proceed for training in step 4 or go back and redo validation with different parameter settings.

Step 4 of 5: Training Algorithm Outputs The results of model building with PLS are displayed in the view.

Step 5 of 5: Class Prediction You can edit the Name and Notes. The dialog informs about the Creation and Last modified date, Owner, Technology, Algorithm Name, Overall Accuracy, Endpoint Name, Number of Endpoints, and Endpoint Value List. Clicking *Finish* will add a node called 'Partial Least Squares Discrimination Model' in the experiment navigator and exit the wizard.

Parameter	Additional Information
Number of Components	Number of components to decompose to; default value is 4.
Scaling	<p>Auto Scaling: Select the Auto Scaling option from the Scaling drop-down list.</p> <ol style="list-style-type: none"> Subtracts the mean μ_i from each m_{ij}: $m_{ij} = m_{ij} - \mu_i.$ Scales down the value by a factor equal to the standard deviation σ_i: $m_{ij} = \frac{(m_{ij} - \mu_i)}{\sigma_i}.$ <p>Pareto: Select the Pareto Scaling option from the Scaling drop-down list. It scales down the value by a factor equal to the square root of the standard deviation σ_i: $m_{ij} = \frac{m_{ij}}{\sqrt{\sigma_i}}.$ </p> <p>No Scaling: You can select No Scaling option to skip scaling.</p>
Validation Type	Only N-Fold validation is supported.
Number of Folds	Sets the number of folds; default value of 3 folds is good to go with.
Number of Repeats	Sets the number of repeats; default value of 10 is good to go with.

Table 21.2: Validation Parameters

21.9 Viewing Classification Results

The results of classification consist of the following views - The [Classification Report](#), and if Class Labels are present in this dataset, the [Confusion Matrix](#) and the [Lorenz Curve](#) as well. These views provide an intuitive feel for the results of classification, help to understand the strengths and weaknesses of models, and can be used to tune the model for a particular problem. For example, a classification model may be required to work very accurately for one class, while allowing a greater degree of error on another class. The graphical views help tweak the model parameters to achieve this.

21.9.1 Confusion Matrix

A Confusion Matrix presents results of classification algorithms, along with the input parameters. It is common to all classification algorithms in **GeneSpring GX** - [classification.SVM](#), [Neural Network](#), [Naive Bayesian Classifier](#), and [Decision Tree](#), appears as follows:

The Confusion Matrix is a table with the true class in rows and the predicted class in columns. The

Confusion Matrix							☐
	[0hr] (Predi...	[1hr] (Predi...	[2hr] (Predi...	[4hr] (Pred...	[8hr] (Predi...	Accuracy	
(True) [0hr]	3	1	0	0	0	75.000	
(True) [1hr]	1	3	0	0	0	75.000	
(True) [2hr]	0	1	3	0	0	75.000	
(True) [4hr]	1	0	0	2	1	50.000	
(True) [8hr]	0	0	0	1	3	75.000	
Overall Accuracy						70.000	

Figure 21.12: Confusion Matrix for Training with Decision Tree

diagonal elements represent correctly classified experiments, and cross diagonal elements represent misclassified experiments. The table also shows the learning accuracy of the model as the percentage of correctly classified experiments in a given class divided by the total number of experiments in that class. The average accuracy of the model is also given. See Figure 21.12

- For [validation](#), the output shows a cumulative Confusion Matrix, which is the sum of confusion matrices for individual runs of the learning algorithm.
- For [training](#), the output shows a Confusion Matrix of the experiments using the model that has been learnt.
- For [classification](#), a Confusion Matrix is produced after classification with the learnt model only if class labels are present in the input data.

21.9.2 Classification Report

This report presents the results of classification. It is common to the three classification algorithms - [Support Vector Machine](#), [Neural Network](#), and [Decision Tree](#).

The report table gives the identifiers; the true Class Labels (if they exist), the predicted Class Labels and class belongingness measure. The class belongingness measure represents the strength of the prediction of belonging to the particular class. See Figure 21.13

21.9.3 Lorenz Curve

Predictive classification in **GeneSpring GX** is accompanied by a class belongingness measure, which ranges from 0 to 1. The Lorenz Curve is used to visualize the ordering of this measure for a particular class.

Identifier	celine	Predicted(celine)	Confidence Measure(C...
MPRO_0hr_A.CEL	[A]	[A]	1.000
MPRO_1hr_A.CEL	[A]	[A]	1.000
MPRO_2hr_A.CEL	[A]	[A]	1.000
MPRO_4hr_A.CEL	[A]	[A]	1.000
MPRO_8hr_A.CEL	[A]	[B]	0.812
MPRO_0hr_B.CEL	[B]	[A]	1.000
MPRO_1hr_B.CEL	[B]	[B]	0.532
MPRO_2hr_B.CEL	[B]	[B]	0.922
MPRO_4hr_B.CEL	[B]	[B]	1.000
MPRO_8hr_B.CEL	[B]	[B]	1.000
MPRO_0hr_C.CEL	[C]	[A]	1.000
MPRO_1hr_C.CEL	[C]	[A]	1.000
MPRO_2hr_C.CEL	[C]	[C]	0.868
MPRO_4hr_C.CEL	[C]	[C]	1.000
MPRO_8hr_C.CEL	[C]	[C]	0.815
MPRO_0hr_D.CEL	[D]	[D]	1.000
MPRO_1hr_D.CEL	[D]	[D]	1.000
MPRO_2hr_D.CEL	[D]	[D]	1.000
MPRO_4hr_D.CEL	[D]	[D]	1.000
MPRO_8hr_D.CEL	[D]	[B]	1.000

Figure 21.13: Decision Tree Classification Report

The items are ordered with the predicted class being sorted from 1 to 0 and the other classes being sorted from 0 to 1 for each class. The Lorenz Curve plots the fraction of items of a particular class encountered (Y-axis) against the total item count (X-axis).

For a given class, the following intercepts on the X-axis have particular significance:

The light red traces the number of items predicted to belong to the selected class.

Classification Quality The point where the red curve reaches its maximum value (Y=1) indicates the number of items which would be predicted to be in a particular selected class if all the items actually belonging to this class need to be classified correctly.

Consider a dataset with two classes A and B. All points are sorted in decreasing order of their belongingness to A. The fraction of items classified as A is plotted against the number of items, as all points in the sort are traversed. The deviation of the curve from the ideal indicates the quality of classification. An ideal classifier would get all points in A first (linear slope to 1) followed by all items in B (flat thereafter). The Lorenz Curve thus provides further insight into the classification results produced by **GeneSpring GX**. The main advantage of this curve is that in situations where the overall classification accuracy is not very high, one may still be able to correctly classify a certain fraction of the items in a class with very few false positives; the Lorenz Curve allows visual identification of this fraction (essentially the point where the red line starts departing substantially from the steady slope line to Y=1). See Figure 21.14

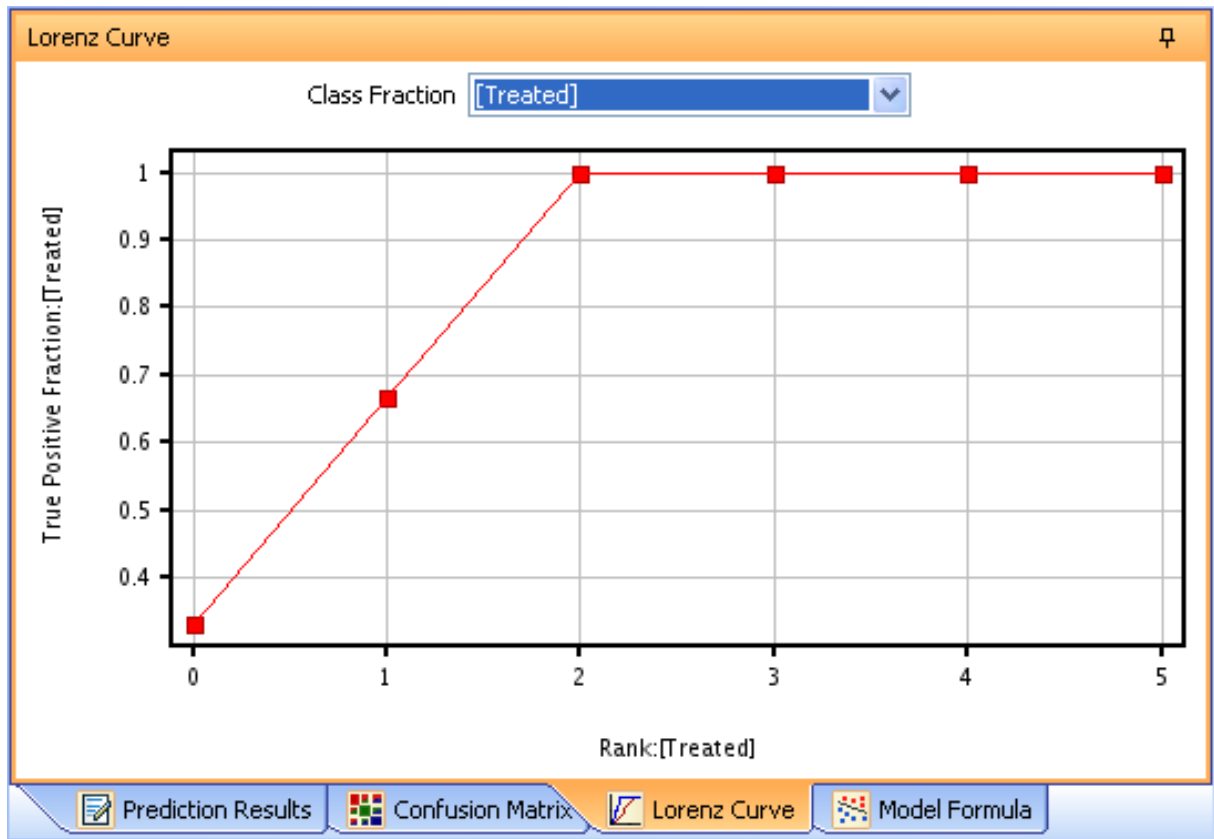


Figure 21.14: Lorenz Curve for Neural Network Training

Lorenz Curve Operations

The Lorenz Curve view is a lassoed view and is synchronized with all other lassoed views open in the desktop. It supports all selection and zoom operations like the scatter plot.

Chapter 22

Gene Ontology Analysis

22.1 Working with Gene Ontology Terms

The Gene Ontology™(GO) Consortium maintains a database of controlled vocabularies for the description of molecular functions, biological processes and cellular components of gene products. These GO terms are represented as a Directed Acyclic Graph (DAG) structure. Detailed documentation for the GO is available at the Gene Ontology homepage (<http://geneontology.org>). A gene product can have one or more molecular functions, be used in one or more biological processes, and may be associated with one or more cellular components. The DAG structure ensures that a gene with a particular GO term also has several other ancestor GO terms implicitly.

In **GeneSpring GX**, the technology associated with an experiment provides GO terms associated with the entities in the experiment. For Affymetrix, Agilent and Illumina technologies, GO terms are packaged with **GeneSpring GX**. For custom technologies, GO terms must be imported and marked while creating custom technology for using the GO analysis. For further details, refer to Step 9 of [Technology Creation](#) in the generic chapters.

GeneSpring GX is packaged with the GO terms and their DAG relationships as provided by the GO Ontology Consortium on their website (<http://geneontology.org>). These ontology files will be updated periodically and these updates will be available via *Annotations* → *Update Technology Annotations* → *From Agilent Server*. Locate and click on *GOData* when the *Automatic Software Update* window appears. Click *Update* button. It is necessary to have an active internet connection to avail this feature.

Users can also update the ontology files directly from GO consortium website using a script. In order to execute the script, do the following:

- Download the OBO file from GO consortium, at (<http://geneontology.org/GO.downloads.ontology.shtml>).

- Open the Script Editor in **GeneSpring GX** from *Tools* → *Script Editor*
- Copy the following script in the Script Editor

```
script.marray.gobrowser.createGOData.writeGOData('godata.bin','gene_ontology_edit.obo')
```
- The first argument in the script (godata.bin) is the output file name. A file by this name containing GO data is prepackaged with the tool and is present in the installation folder
Agilent/GeneSpringGX/app/DataLibrary/GOData/GeneOntologyData/.
Take a back up of this file elsewhere if you want to retain the last update information. On running the script, the new updates will be saved as godata.bin in the folder
Agilent/GeneSpringGX/bin/launcher/lib/ by default. Either move this update file to the folder
Agilent/GeneSpringGX/app/DataLibrary/GOData/GeneOntologyData/ after running the script or specify this folder in the script before running it. Note that the godata.bin needs to be present at the location
Agilent/GeneSpringGX/app/DataLibrary/GOData/GeneOntologyData/ for doing the GO analysis.
- The second argument (gene_ontology_edit.obo) is the input OBO file downloaded by the user. In the script, provide the correct name of the file and its path and then run it. The updates get saved as godata.bin in the folder specified in the script or in Agilent/GeneSpringGX/bin/launcher/lib/ by default.

Custom GO annotation file (mapping of probe-ID to GO terms) from any source can also be imported in **GeneSpring GX**. This can be done while creating a generic single or two color technology, or while updating an existing technology using *Annotations* → *Update Technology Annotations*. For carrying out GO analysis, the custom annotation file can either contain a single column with all the GO IDs in it, separated by a separator or it can contain separate columns for the different GO processes. Some of the GO formats supported by **GeneSpring GX** is given below (not inclusive of all):

- GO:0000012
- go:012
- 12
- GO:0000012(single strand break repair);GO:0000910 (cytokinesis);GO:0006260 (DNA replication);GO:000626 (DNA ligation);GO:0006281(DNA repair);GO:0006310 (DNA recombination);GO:0008150 (biological_process)

In case of multiple columns, while each column can be in a different format, multiple formats within a column is not supported.

The single column with multiple GO IDs should be marked as *Gene Ontology accession* from the dropdown menu. Instead if columns containing individual GO processes(Biological Process, Cellular Component and Molecular Function) are present, they should be marked accordingly in the dropdown menu.

22.2 Introduction to GO Analysis in GeneSpring GX

GeneSpring GX has a fully-featured gene ontology analysis module that allows exploring gene ontology terms associated with the entities of interest. **GeneSpring GX** allows the user to visualize and query the *GO Tree* dynamically, to view GO terms at any level as a *Pie Chart*, to dynamically drill into the pie, to navigate through different levels of the GO tree, to compute enrichment scores for GO terms based upon a set of selected entities, and to use enrichment scores and FDR corrected p-values to filter the selected set of entities. The results of GO analysis can then provide insights into the biology of the system being studied.

In the normal flow of gene expression analysis, GO analysis is performed after identifying a set of entities that are of interest, either from statistical tests or from already identified gene lists. You can select a set of entities in the dataset and launch GO analysis from the *results Interpretation* section on the workflow panel.

Note: To perform GO Analysis, GO terms associated with the entities should be available. These are derived from the technology of the experiment. For Affymetrix, Agilent and Illumina technologies, **GeneSpring GX** packages the GO Terms associated with the entities. For custom technologies, GO terms must be imported and marked while creating custom technology for using the GO analysis.

This chapter details GO Analysis, the algorithms to compute enrichment scores, the different views launched by the GO analysis and methods to explore the results of GO analysis.

22.3 GO Analysis

GO Analysis can be accessed from most of the workflows in **GeneSpring GX**. Clicking on the *GO Analysis* link in the *Results Interpretations* section on the workflow panel will launch a wizard that will guide you through collecting the inputs for the analysis and creating an entity list with the significant entities.

Input Parameters The input parameter for GO analysis is any entity list in the current active experiment. By default, the active entity list in the current experiment is shown as the chosen entity list. Clicking on *Choose* will show a tree of entity lists in the current experiment. You can choose any of the entity lists and launch GO Analysis. See Figure [22.1](#)

Output Views The results of GO Analysis are shown in the view. Depending upon the input entity list, GO terms that are enriched with a p-value cut-off of 0.1 are shown. If no entities satisfy the cut-off, click on the *Change cutoff* button and change the cut-off from the slider or in the text box. This will dynamically update the views accordingly

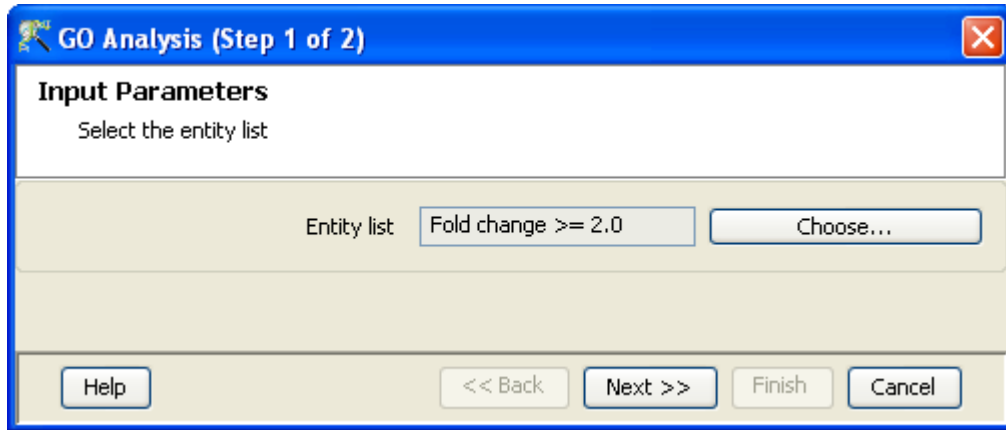


Figure 22.1: Input Parameters

The output views show a [pie chart](#), a [spreadsheet](#) with the GO terms that satisfy the p-value cut-off and a [GO Tree](#). These GO terms have been derived from only those GO IDs that were part of the input entity list and obeying the cut-off p-value. The spreadsheet has GO accession numbers, their description, uncorrected p-values obtained from the hypergeometric test (described later), Benjamini-Yekutieli corrected p-values to take care of multiple GO term testing and correlation between them (the p-value cutoff is based on the corrected values), along with frequencies of occurrence of GO terms in the input entity list and in the All Entities list.

All the views are interactive and are dynamically linked. Thus clicking on the pie chart will select the GO Term in the GO tree and will show the corresponding entities associated with the GO terms. Clicking on a GO term on the spreadsheet will highlight the corresponding term in the GO Tree and show the corresponding entities. For details on the views and navigation see the section on [GO Analysis Views](#). See [Figure 22.2](#)

Click *Finish* to save the entity lists corresponding to relevant GO terms. This will create the appropriate entity lists in a folder called GO Analysis. You can also manually select a set of GO terms and save entity lists corresponding these GO terms using the Save Custom button.

The p-value for each GO term reflects the enrichment in frequency of that GO term in the input entity list relative to the All Entities list. The p-value for a GO term g is determined by the following quantities:

- Number of entities in the input entity list which have the term g or any descendant term.
- The number of entities in the All Entities list which have the term g or any descendant term.
- The total number of entities in the input entity list, and
- The total number of entities in the All Entities list.

Note that **GeneSpring GX** takes GO terms from Biological Processes, Molecular functions and Cellular components together. For details on the computation of the enrichment score or p-value [see below](#).

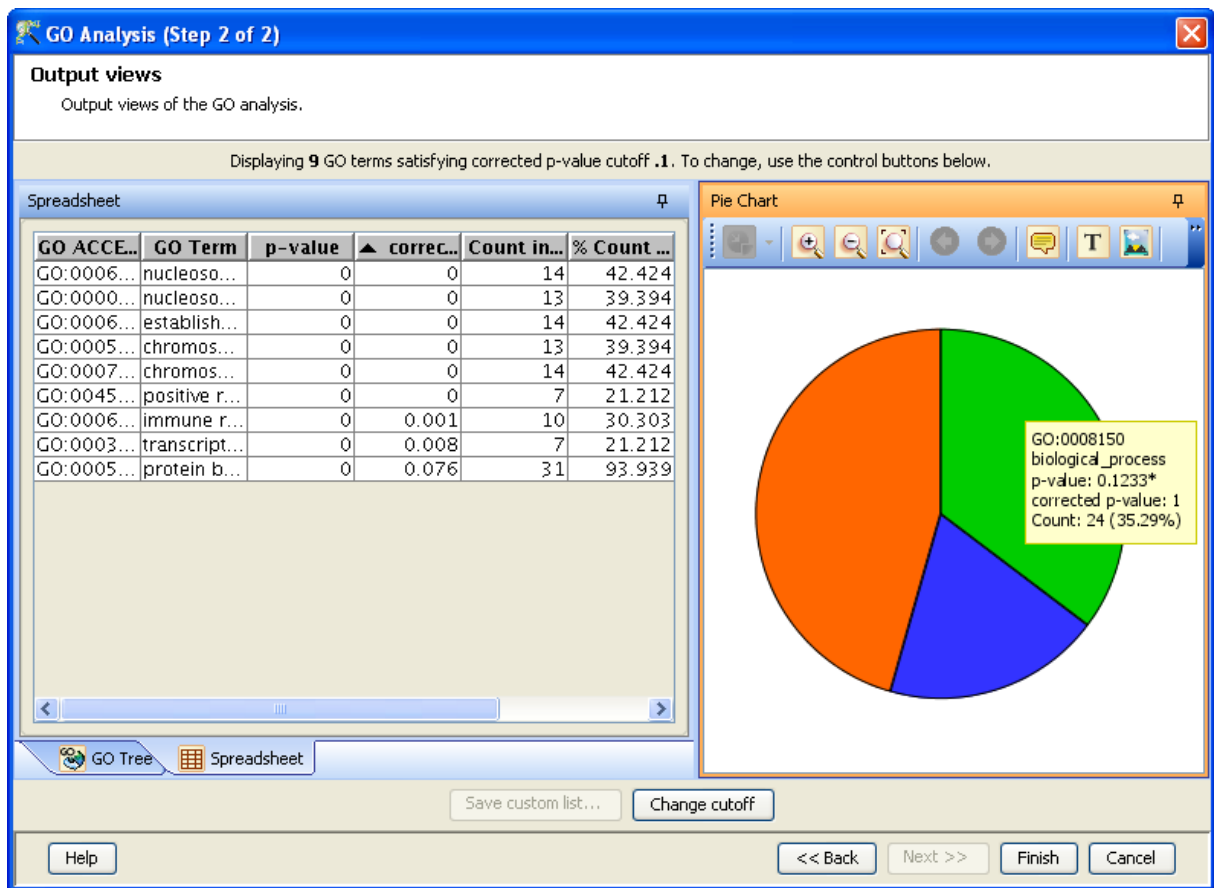


Figure 22.2: Output Views of GO Analysis

Note the following rules govern which GO terms are displayed in the various output views.

- All GO terms associated with the input entity list are taken and filtered by the given p-value cutoff.
- For each term that survives, all ancestors which pass the p-value cut off are considered as well
- Terms from 1 and 2 above are shown in the output table view. So it could happen that a term g has p-value .09, say, but it does not show up at cut off .1 because there is no gene in the input list that is directly associated with g ; there are genes associated with descendants of g but those terms do not satisfy the p-value cut off; at a higher cut off those terms could show up and they drag g in.
- The **Count in Selection** shown in the table for a GO term g is obtained as follows: take all descendant GO terms which pass the cut-off and add up genes in the input list associated with these terms. Add to this genes in the input list associated with g itself. This is the count in selection. Note that this count could change as the cutoff changes.
- When saving entity lists for GO terms, what is saved are exactly those entities which contribute to the count above.
- The **Count in Total** is obtained as follows: just add up all genes in All Entities associated with this term or any descendant term (independent of cut-offs).

22.4 GO Analysis Views

22.4.1 GO Spreadsheet

The GO Spreadsheet shows the following for each Go Term displayed.

p-value: The probability of obtaining the specified GO accession number from a list of random entities. Less the p-value more significant is the GO accession number.

corrected p-value: Since multiple GO accession number are tested for their significance, hence a multiple testing correction is performed. The GO spreadsheet is sorted based on the corrected p-values.

Count in Selection: This refers to the number of genes in the selected entity (for example, from T-test) list which have that particular GO term.

% Count in Selection: This refers to the percentage of genes in the input entity list which have that GO term.

Count in Total: This refers to the number of genes in All Entities which have that GO term.

% Count in Total: This refers to the percentage of genes in the All Entities list which have that GO term.

The Selection of GO terms in this table will select the corresponding GO terms in the *GO Tree* view and will show the entities associated with the GO term. See Figure [22.3](#)

22.4.2 The GO Tree View

The *GO Tree* view is a tree representation of the GO Directed Acyclic Graph (DAG) as a tree view with all GO Terms and their children. Thus there could be GO terms that occur along multiple paths of the GO tree. The GO tree is represented on the left panel of the view. The panel to the right of the GO tree shows the list of entities in the experiment that corresponds to the selected GO term(s). The selection operation is detailed below. See Figure [22.4](#)

The GO tree is always launched expanded up to three levels. The GO tree shows the GO terms along with their enrichment p-value in brackets. The GO tree shows only those GO terms along with their full path that satisfy the specified p-value cut-off. GO terms that satisfy the specified p-value cut-off are shown in blue, while others that are on the path and do not satisfy the cut-off are shown in black.

Note that the final leaf node along any path will always have GO term with a p-value that is below the specified cut-off and shown in blue. Also note that along an extended path of the tree there could be multiple GO terms that satisfy the p-value cut-off.

GO ACCE...	GO Term	p-value	▲ correc...	Count in...	% Count ...	Count in...
GO:0006...	nucleoso...	0	0	14	42.424	82
GO:0000...	nucleoso...	0	0	13	39.394	80
GO:0006...	establish...	0	0	14	42.424	210
GO:0005...	chromos...	0	0	13	39.394	151
GO:0007...	chromos...	0	0	14	42.424	248
GO:0045...	positive r...	0	0	7	21.212	15
GO:0006...	immune r...	0	0.001	10	30.303	344
GO:0003...	transcript...	0	0.008	7	21.212	126
GO:0005...	protein b...	0	0.076	31	93.939	4039

Figure 22.3: Spreadsheet view of GO Terms.

The *GO Tree* provides a link between the GO terms and the entities in the experiment. Operations on the GO Tree are detailed below:

Expand and Collapse the GO tree: The GO tree can be expanded or collapsed by clicking on the root nodes.

GO Tree Labels: The GO tree is labelled with GO terms as default. You can change the GO tree to be labelled by either the GO Accession; the GO terms; or both from the right-click properties dialog.

p-value and Count: The number in the bracket corresponding to a GO term shows the p-value or enrichment value of the GO term. You can display the p-value, the actual counts of both the p-value and the actual counts for the GO term from the right-click properties dialog.

The counts show two values. The first value shows the number of entities in the entity list contributing to any significant GO term in the hierarchy. The second count value shows the number of entities that contribute any significant GO term in the hierarchy in the experiment.

Select Genes: Clicking on a GO term in the tree will select the entities in the entity list that contributed to any significant GO term in the hierarchy.

You can choose multiple GO terms in the tree and see *All Genes* that contributed to any significant GO term in the hierarchies. This will show a union of all the entities corresponding to the selected GO terms. Or you can choose multiple GO terms in the tree and select the *Common Genes* that contributed to any significant GO term in the hierarchies. This will show an intersection of the entities corresponding to the selected GO terms. See Figure 22.5

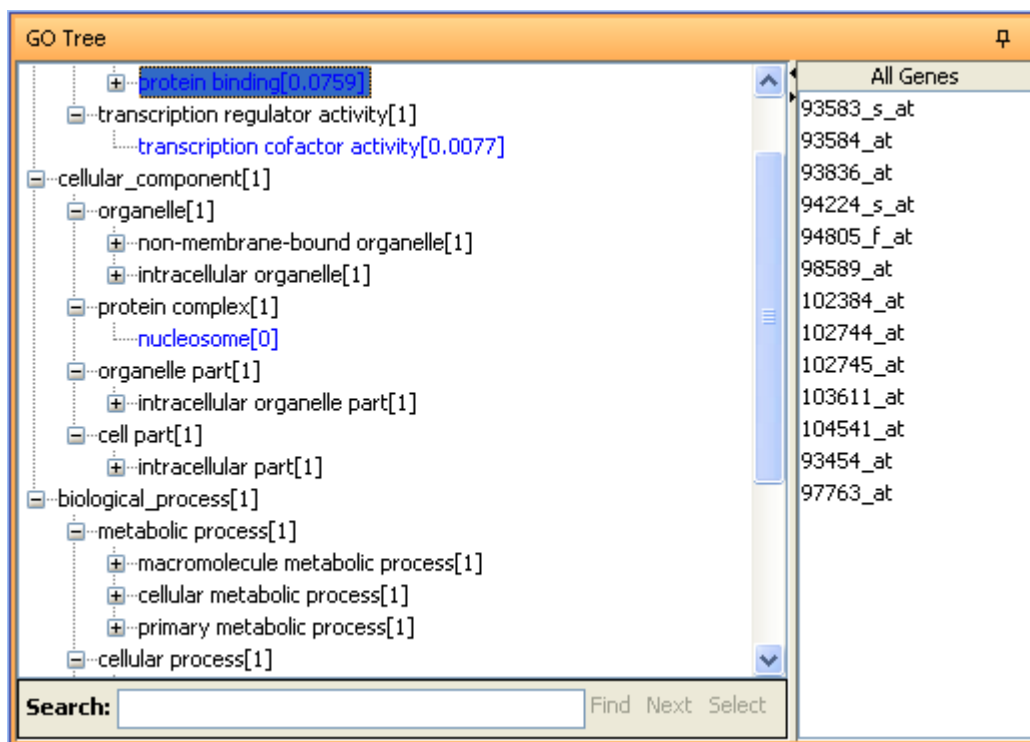


Figure 22.4: The GO Tree View.

Selecting *Show All Genes* or *Show Common Genes* can be chosen from the right-click *Properties* menu of the GO tree.

22.4.3 The Pie Chart

The pie chart view shows a pie of the GO terms with the number of entities that contribute to the any significant GO term in the hierarchy. When the pie chart is launched, it is launched with the top level GO terms of *Molecular Function*, *Biological Process* and *Cellular Component*. The slices of the pie is drawn with the number of entities in each of the three terms that contribute to any significant GO terms in whole hierarchy of GO terms. See Figure 22.6

The pie chart view is rich with functionality. It allows you to drill into the pie and reach any level of the GO tree, and navigate through the different drill levels. You can select the entities corresponding to the pies or the GO terms in any view. The pie chart allows you to zoom in and out of view, fit the pie chart to view, enable and delete callouts for the slices, add text and images to the view and create publication quality outputs. The functionality of the pie chart is detailed below:

Default launch The pie chart by default is launched with the three top level GO terms of *Molecular Function*, *Biological Process* and *Cellular Component*.

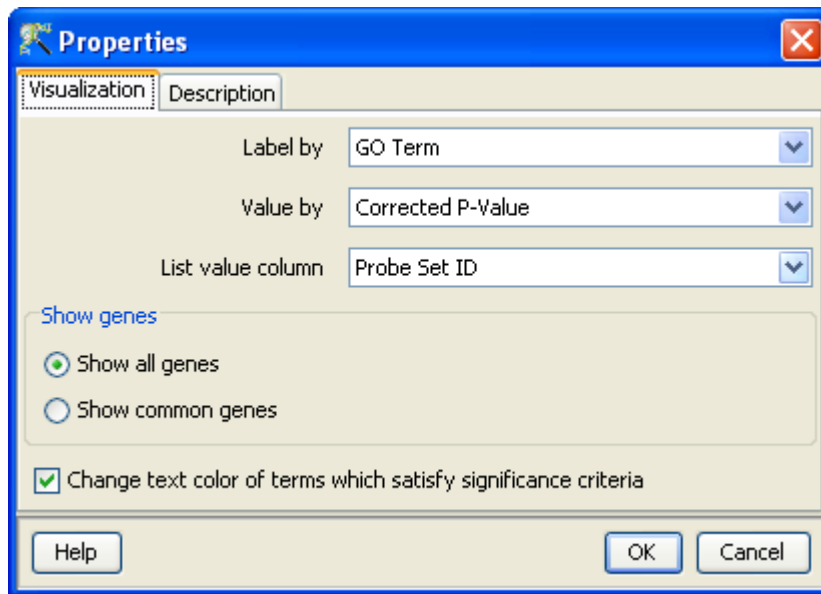


Figure 22.5: Properties of GO Tree View.

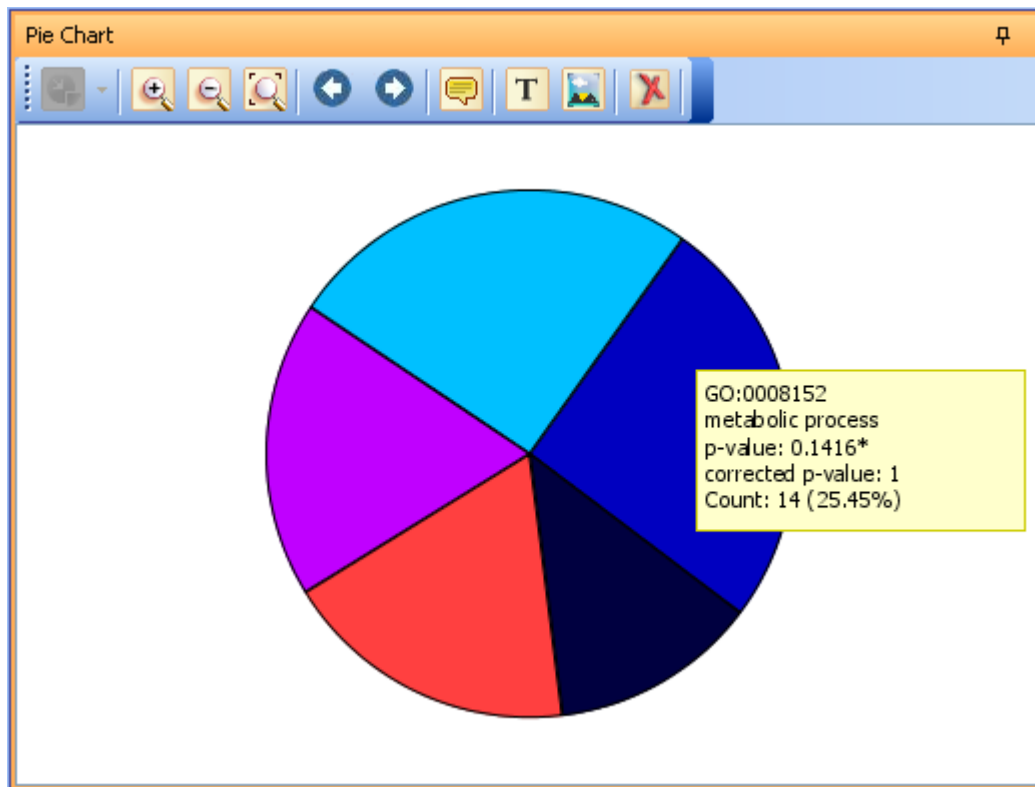




Figure 22.6: Pie Chart View.

Selecting Slices of the Pie To select a slice of a pie, click on the slice of interest. To add to the selection *Shift + Left-click* on the pies of interest. All the selected pies will be shown with a yellow border. You can also select slices by clicking and dragging the mouse over the canvas. A selection rectangle will be shown and all the slices within the selection rectangle will be selected.




Drill into pie To drill into a GO term and traverse down the hierarchy, select the pie or pies of interest by clicking on it. Click the Drill Selected Pie  icon on the toolbar. This will execute one of the four selected options that are chosen in the drop-down list of the Drill Selected Pie  icon. Double-click on any pie has exactly the same effect as drilling down the slice according to the chosen option.



Drill Pie One-Level This option will replace the current pie chart with a new pie chart, with GO terms one level below the GO terms of the selected slices. For example, if *Molecular Function* is selected, and the *Drill Pie One-Level* option is chosen, then the current top level pie will be replaced a pie with the first level children of *Molecular Function*. This is the default option.

Drill Pie All-Levels This option will replace the current pie chart with a new pie chart, with all the GO terms of the selected slices(s) below the GO terms of the selected slice(s). This pie chart cannot be drilled down further since it has been expanded to the last level.

Expand Slice One-Level This option will expand the selected slice(s) with GO terms one level below the GO terms of the selected slices. The other unselected slices, their GO terms, and their counts will remain unaffected. However, the slice sectors may change depending upon the counts of the individual slices


Expand Slice All-Levels This option will expand the selected slice(s) with all the GO terms of the selected slice(s) below the GO term of the selected slice(s). The other unselected slices, their GO terms, and their counts will remain unaffected. However, the slice sectors may change depending upon the counts of the individual slices

Zoom and fit to view To zoom in, zoom out or fit the pie chart view to the displayed canvas, click on the zoom in  icon zoom out  icon and Fit to view  icon respectively.

Navigating through pies In the course of exploring the GO Analysis pie chart, you may be drilled into different levels of selected slices using different drill methods detailed above. You can navigate between the different drilled states of the pie chart by clicking on the Back  icon and Forward  icon respectively. These icons will be enabled or disabled appropriately depending upon the current state of the pie chart.

The pie chart can only remember a single path from the original top level pie to the current state. Thus, for example, if you had drilled into one slice, then went back, choose another slice to drill into then the previous drilled path will not be maintained.

Callouts for slices The slices of the pie chart denote different GO terms. If you hover the mouse on the slice the tool-tip shows the associated GO ID; the GO term; the p-value of the GO term; and the count of the number of entities contributing to any significant GO term in the hierarchy. Note that GO terms could be present even if they did not pass the specified cut-off because a GO term that was lower in the hierarchy satisfied the p-value cut-off. We use an asterisk (*) in the p-value to indicate this.

You can create a callout for selected slices by selecting the slices of interest and clicking on the Show Callouts  icon on the tool bar. This will create a callout with the GO ID; the GO term; the p-value of the GO term; and the count of the number of entities contributing to any significant GO

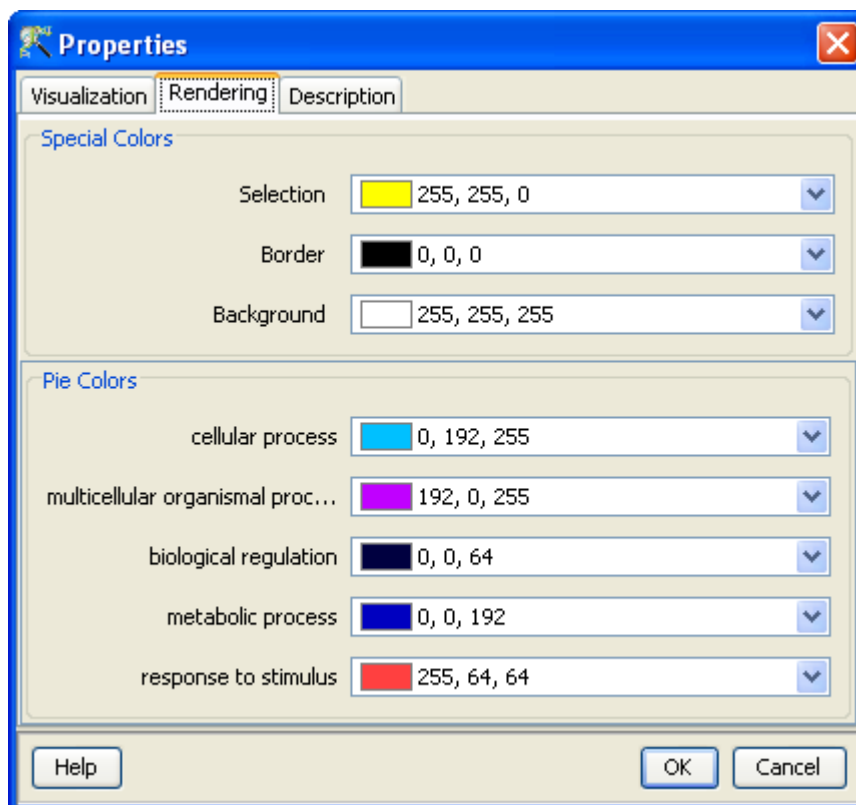





Figure 22.7: Pie Chart Properties.

term in the hierarchy. The callouts can be selected, moved, and resized. To delete a callout, select the callout and click the Delete  icon icon.

Add text and Image Texts can be added to the pie chart wherever required. To add text to the pie chart, click on the Switch Text Mode  icon. This will change the cursor. You can click on the canvas of the pie chart and add text. Click on the icon again to toggle back to the selection mode. To add an image to the pie chart, click on the Insert Image  icon. This will pop-up a file chooser. Choose the required image and add it to the pie chart.

Right-click menu on the pie chart The right click menu on the pie chart has options to print the pie chart to a browser, export the pie chart as an image to any desired resolution; and access the properties of the pie chart. The properties options of the pie chart allows you to change the properties of the view as detailed below: See Figure 22.7

Visualization The *Visualization* tab of the properties dialog allow you to change the height of the pie chart from 0 to 100. the default is set at 100, when the pie chart is represented as a circle. The height can be decreased to make the pie chart an ellipse.

The *Minimum row count* of the pie chart can be changed. The default is set to 1. If the count, or number of entities is less than that specified in this dialog, the slice will not be displayed. This can be used to filter out GO terms with only a small number of entities.

Rendering The selection color, the border color, the background color, and the color of the slices of the pie can be changed.

Description You can add any description to the pie chart from the *Description* tab.

22.5 GO Enrichment Score Computation

Suppose we have selected a subset of significant entities from a larger set and we want to classify these entities according to their ontological category. The aim is to see which ontological categories are important with respect to the significant entities. Are these the categories with the maximum number of significant entities, or are these the categories with maximum enrichment? Formally stated, consider a particular GO term G . Suppose we start with an array of n entities, m of which have this GO term G . We then identify x of the n entities as being significant, via a t-test, for instance. Suppose y of these x entities have GO term G . The question now is whether there is enrichment for G , i.e., is y/x significantly larger than m/n . How do we measure this significance?

In most arrays each probeset is associated with single or multiple GO terms. Since some genes (Entrez-ids) are represented by multiple probesets, therefore GO term enrichment calculation gets biased toward genes having multiple probesets. Hence for unbiased calculation, multiple probesets corresponding to the same Entrez id are collapsed before running the GO analysis. A union of GO terms corresponding to multiple probesets for the same Entrez id is used for collapsed probeset. The following rule sets are followed for systematically condensing the probesets:

- If the entity has a single Entrez ID then take associated GO terms and associate it with this Entrez ID.
- If an entity has multiple Entrez IDs then if the Entrez ID has occurred previously and has an associated GO term, these are removed from the list. Each remaining Entrez ID get is then associated with GO terms.

GeneSpring GX computes a p-value to quantify the above significance. This p-value is the probability that a random subset of x entities drawn from the total set of n entities will have y or more entities containing the GO term G . This probability is described by a standard hypergeometric distribution (given n balls, m white, $n-m$ black, choose x balls at random, what is the probability of getting y or more white balls). **GeneSpring GX** uses the hypergeometric formula from first principles to compute this probability.

Since very often large number of hypothesis will be tested, some form of correction is required. However, there is no simple or straight forward way to do that. The different hypotheses are not independent by virtue of the way that GO is structured and even with this difficulty addressed, we are most interested in patterns of p-values that correspond to a structure in GO rather than single p-values exceeding some fixed threshold. In **GeneSpring GX** we have addressed the first issue using the Benjamini-Yekutieli correction [53, 4], which takes into account the dependency among the GO terms.

Finally, one interprets the p-value as follows. A small p-value means that a random subset is unlikely to match the actually observed incidence rate y/x of GO term G , amongst the x significant entities.

Consequently, a low p-value implies that G is enriched (relative to a random subset of x entities) in the set of x significant entities.

NOTE: The GO analysis implementation in **GeneSpring GX** considers Molecular Function, and Cellular Component all together. Further, “part-of” relations in the ontology are ignored.

Chapter 23

Gene Set Enrichment Analysis

23.1 Introduction to GSEA

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant differences between two phenotypes. Traditional analysis of expression profiles in a microarray experiment involves applying statistical analysis to identify genes that are differentially expressed. In many cases, few genes pass the statistical significance criterion. When a larger number of genes qualify, there is often a lack of unifying biological theme, which makes the biological interpretation difficult. GSEA overcomes these analytical difficulties by focussing on gene sets rather than individual genes. It uses the ranked gene list to identify the gene sets that are significantly differentially expressed between two phenotypes.

GSEA analysis in **GeneSpring GX** is based on the GSEA implementation by the Broad Institute (<http://www.broad.mit.edu/gsea>) The current chapter details the GSEA Analysis, the algorithms to compute enrichment scores and methods to explore the results of GSEA analysis in **GeneSpring GX** .

23.2 Gene sets

A gene set from the Broad Institute is a group of genes, based on prior biological knowledge, that share a common function, chromosomal location or regulation. In **GeneSpring GX**, gene sets can also be defined as any entity lists created in the application that are used for GSEA.

The Broad Institute (<http://www.broad.mit.edu/index.html>) maintains a collection of gene sets. **GeneSpring GX** supports the import of MIT-Harvard-Broad gene sets in the following file formats:

- **txt/csv:** First line is header information and the remaining lines are genes.

- **grp**: Gene set file format where each gene is in a new line
- **gmt**: Gene Matrix Transposed file format where each row represents a gene set
- **xml**: Molecular signature database file format (msigdb_*.xml)

A detailed description of the file formats can be found at http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats. The Broad gene sets can be found at http://www.broad.mit.edu/gsea/msigdb/msigdb_index.html. Each individual gene set can be viewed, downloaded and imported into **GeneSpring GX**. Alternatively, after registering with the web-site, one can download the entire collection.

Once Broad gene sets have been downloaded, they can be imported into **GeneSpring GX**. To import the Broad gene sets, click on the *Import BROAD GSEA Gene sets* link from the **Tools** section of the menu bar.

Importing gene sets in .grp, .gmt or .xml formats into **GeneSpring GX** converts them into **GeneSpring GX** Gene Lists which are automatically marked as Gene Symbol. (Note that importing the msigdb_v2.xml into **GeneSpring GX** takes around 10 minutes as the XML file is parsed)

Note: To perform GSEA, the Entrez ID or Gene Symbol mark is essential. These are derived from the technology of the experiment. For Affymetrix, Agilent and Illumina technologies, **GeneSpring GX** packages the Entrez ID and Gene Symbol IDs marks. For custom technologies, Entrez ID or Gene Symbol must be imported and marked while creating custom technology for using the GSEA.

23.3 Performing GSEA in GeneSpring GX

GSEA can be accessed from most of the workflows in **GeneSpring GX**. Clicking on the *GSEA* link in the *Result Interpretations* section of the Workflow panel will launch a wizard that will guide you through GSEA in **GeneSpring GX**.

Input Parameters The input parameters for GSEA analysis is an entity list and an interpretation in the current active experiment. By default, the active entity list and the active interpretation in the experiment are selected. Clicking on the *Choose* option will show a tree of entity lists or interpretations in the experiment. You can choose any of the entity lists and interpretation from the tree as inputs to the GSEA Analysis. See Figure 23.1

Pairing Options In the Pairing Options page, you can explicitly select pairs of conditions for GSEA, or, you can select all the conditions in the interpretation against a single control condition. If you choose pairs of conditions, the table shows all the pairs. Choose the pairs of conditions to test by checking

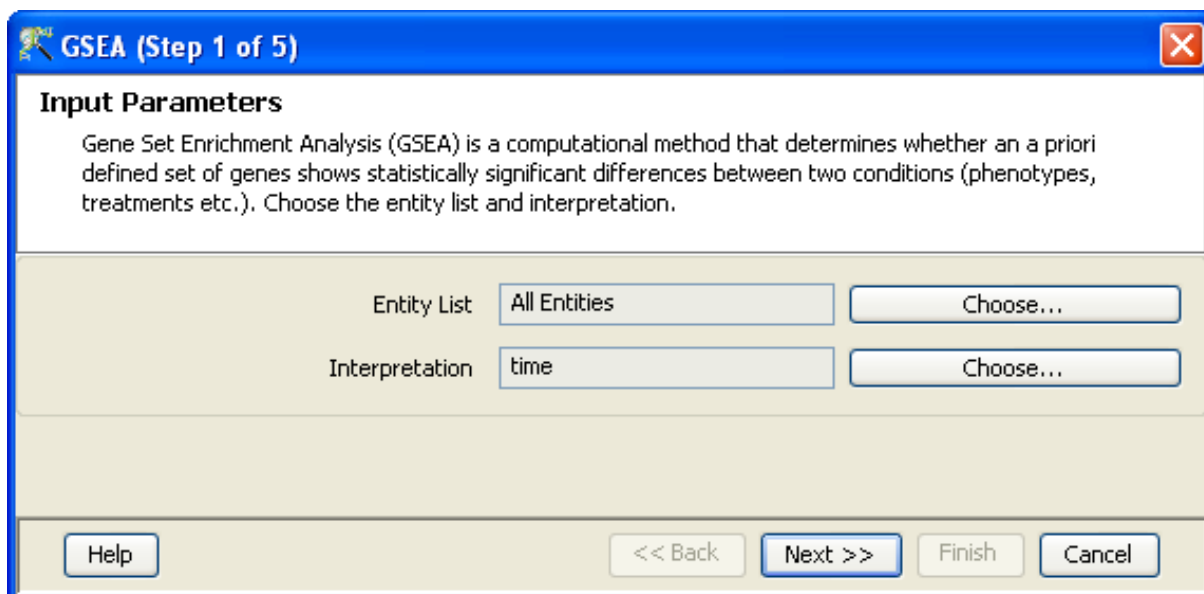


Figure 23.1: Input Parameters

off the corresponding boxes. If you choose all conditions against control, select the condition to use as the control from the drop-down menu. See Figure 23.2

Choose Gene Sets In the *Choose Gene Sets* options page, you can choose one or more of the BROAD gene sets that have been imported. Alternatively, you can select custom gene sets from entity lists that you have created in **GeneSpring GX**. To do this, click on the Advanced Search radio button, search for the entity lists of interest, and select the ones to be used as gene sets for GSEA. See Figure 23.3

You can also specify the minimum number of genes that must match between the gene set and the input entity list for GSEA in order for the gene set to be considered in the analysis. The default is set at 15 genes. Thus, if a gene set has less than 15 genes matching the entity list, then this gene set will not be considered. The default number of permutations used for analysis is set at 100.

Results from GSEA The *Gene Sets satisfying minimum Gene requirement* spreadsheet shows the gene sets with q values below the specified cutoff. The *Gene Sets with fewer than minimum necessary matches* spreadsheet shows the gene sets satisfying the minimum number of matching genes specified in step 3. You can change the q-value cut-off by clicking on the Change q-value cut-off button and entering a new value. See Figure 23.4

GSEA results spreadsheet reports the following columns of values:

- **Gene Sets:** List of gene sets that pass the threshold criterion.
- **Description:** User supplied description associated with the gene set.
- **Total Genes:** Total number of genes in the gene set.

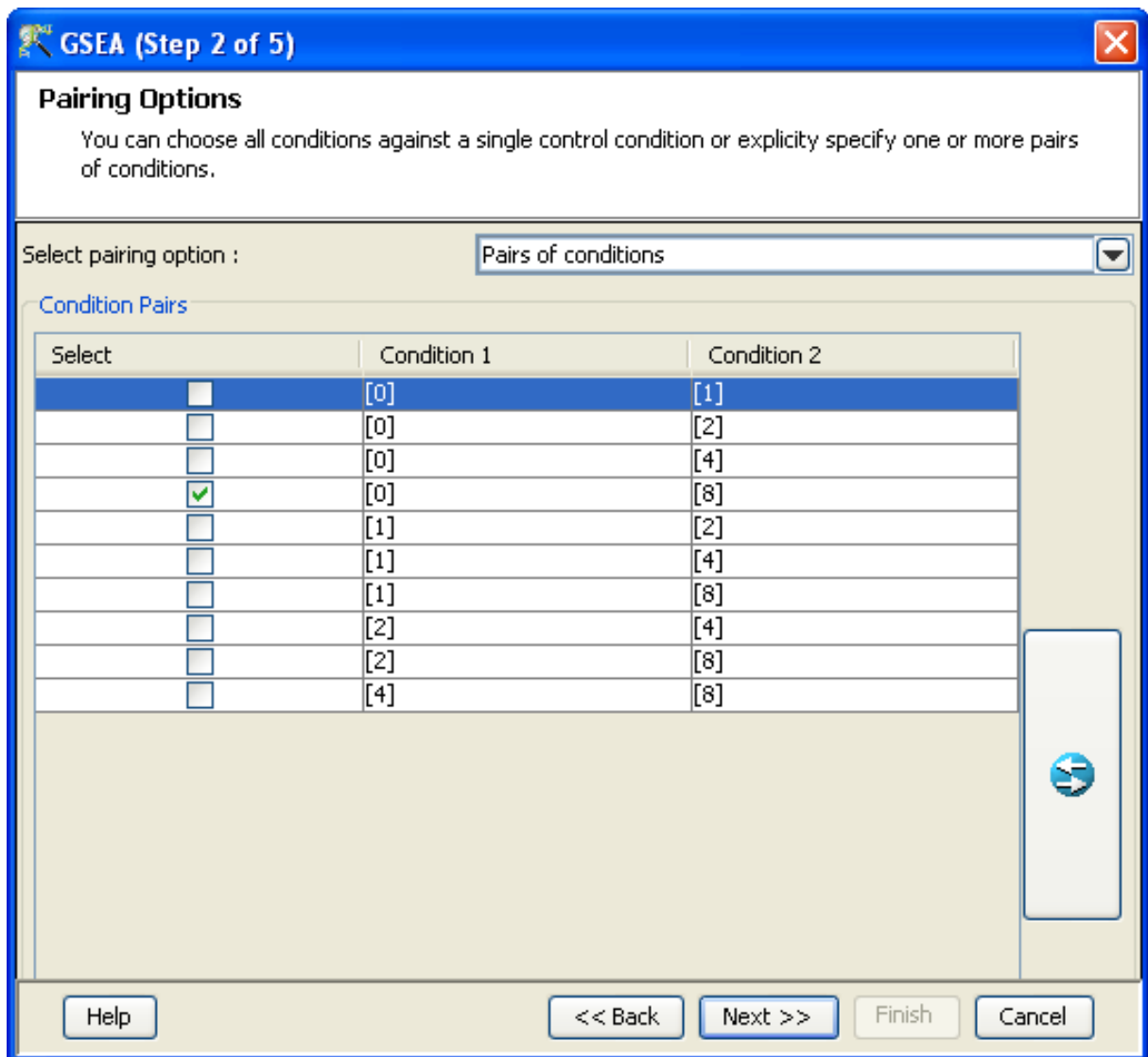


Figure 23.2: Pairing Options

- **Genes Found:** Number of gene in the gene set that are also present in the dataset on which analysis is performed.
- **p value:** Nominal p-value (from null-distribution of the gene-set)
- **q value:** False Discovery Rate q-value
- **ES value:** Enrichment score of the gene set for the indicated pairs of conditions.
- **NES value:** Normalized enrichment score of the gene set for the indicated pairs of conditions.

Last four columns are repeated when multiple pairs of conditions are selected for analysis.

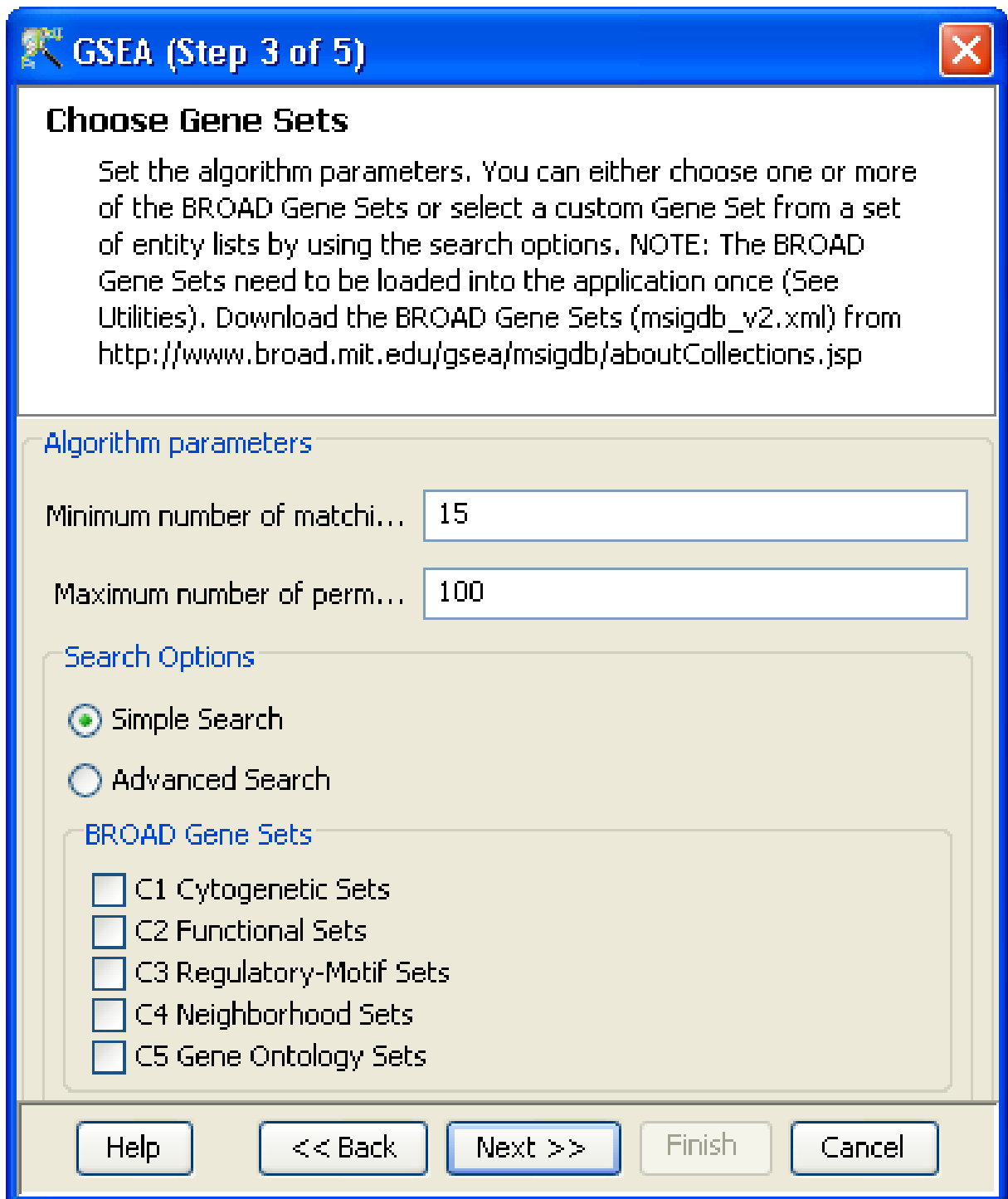


Figure 23.3: Choose Gene Lists

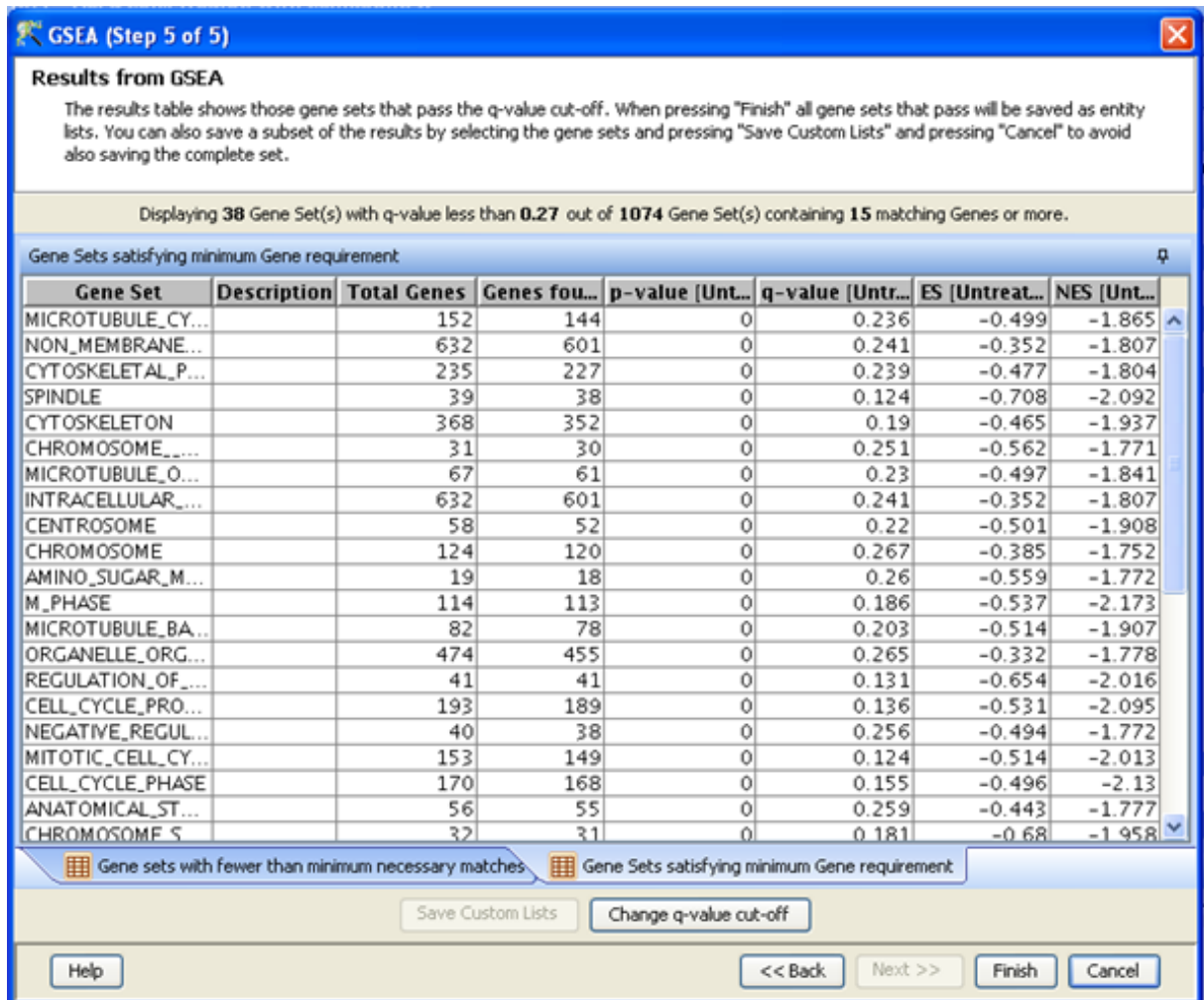


Figure 23.4: Results

Gene sets with q-values below the cutoff can be saved to the Navigator. Click *Finish* to save all the gene sets within the *Gene Sets satisfying minimum Gene requirement* spreadsheet. To save a subset of these gene sets, select the gene sets of interest and click Save Custom Lists. These gene sets will be automatically translated to the technology of the experiment and saved as entity lists in a GSEA folder within the Navigator. The saved entity lists are named according their respective gene set names.

23.4 GSEA Computation

GSEA analysis works on a ranked list of genes to compute the enrichment scores for gene sets. Correlation value for each gene is computed as the ratio between difference in mean expression values and the sum of standard deviations within each phenotype. Correlation values is used rank the genes in the dataset. Enrichment score for a gene set is defined as the maximum deviation from zero, when walking through the ranked list of genes, incrementing the running score for a hit, and decrementing for a miss. A hit is

observed when the gene in the ranked list is present in the gene set, and a miss otherwise.

Thus analysis is restricted to log summarized datasets. If a gene has multiple probes in the dataset, the probe with maximum inter quartile expression range value is considered to compute the mean. Inter quartile range is immune to baseline transformation and hence GSEA results on baseline transformed data and no baseline transformed data remains same. GSEA algorithm and computation of associated metric is detailed in the paper http://www.broad.mit.edu/gsea/doc/gsea_pnas_2005.pdf. The permutative procedure described in the paper is used to compute the *p-values* and *q-values*. Number of permutations can be configured from *Tools* → *Options* → *Data Analysis Algorithms* → *GSEA* of the menu bar.

Pseudo code for GSEA in **GeneSpring GX**

Create a collapsed with rows with unique genes for a given pair of phenotypes

For each row in the dataset

$$m_1 = \text{mean}(\text{phenotype}_1)$$

$$m_2 = \text{mean}(\text{phenotype}_2)$$

$$\text{stdev}_1 = \text{variance}(\text{phenotype}_1)$$

$$\text{stdev}_2 = \text{variance}(\text{phenotype}_2)$$

$$\text{correlationValue} = \frac{m_1 - m_2}{\text{Sqrt}(\text{stdev}_1 + \text{stdev}_2)} \quad \text{Note: stdev is forced to a minimum value of } 0.2 * \text{abs}(\text{mean}_i)$$

Sort correlation values in increasing order to get the RankedGeneList

For each gene set

runningScore = 0

R = number of rows in dataset

g = number of genes in gene set

correlationTotal = sum (correlationValues for genes in gene set)

for gene in RankedGeneList

 if gene present in gene set

 runningScore = runningScore + (correlationValue of gene)/(correlationTotal)

 else

 runningScore = runningScore - 1/(R-g)

ES = max(abs(runningScore))

Compute ES_* for each random permutation of phenotype ordering Divide the enrichment scores obtained into $ES_*(\geq 0)$ and $ES_*(< 0)$

If $ES > 0$

$$p\text{-value} = \frac{\text{NumberOfPermutationsWhereES} > ES_* \text{ and } ES_* \geq 0}{\text{NumberOfPermutationsWhereES}_* \geq 0}$$

$$NES = ES / \max(ES_*(> 0))$$

$$NES_* = NES / \max(ES_*(> 0))$$

For a given gene set with NES_i

$$ratio1 = \frac{\text{NumberOfRandomPermutationsForAllGeneSetsWhereNES} < NES_* \text{ and } NES_* \geq 0}{\text{NumberOfRandomPermutationsForAllGeneSetsWhereNES}_* \geq 0}$$

$$ratio2 = \frac{\text{NumberOfGeneSetsWhereNES}_i < NES \text{ and } NES \geq 0}{\text{NumberofGeneSetsWhereNES} > 0}$$

$$q - value = \frac{ratio1}{ratio2}$$

The process is repeated for gene sets with $ES < 0$

23.5 Import BROAD GSEA Genesets

GSEA can be performed using the 5 genesets which are available from the BROAD Institute's website (<http://www.broad.mit.edu/gsea/>). These genesets can be downloaded and imported into the **GeneSpring GX** to perform GSEA. Clicking on this option allows the user to navigate to the appropriate folder where the genesets are stored and select the set of interest. The files should be present either in .xml or .grp or .gmt formats. Remember to provide 'Read' permissions to these gene lists after import, if you intend having others use these lists to run GSEA.

Chapter 24

Gene Set Analysis

24.1 Introduction to GSA

Gene Set Analysis (GSA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant differences between two phenotypes. Traditional analysis of expression profiles in a microarray experiment involves applying statistical analysis to identify genes that are differentially expressed. In many cases, few genes pass the statistical significance criterion. When a larger number of genes qualify, there is often a lack of unifying biological theme, which makes the biological interpretation difficult. GSA overcomes these analytical difficulties by focussing on gene sets rather than individual genes. The implementation uses the S-Score, computed as a cumulative sum of t-Scores across the phenotypes for genes in a gene set for significance testing.

GSA in **GeneSpring GX** is based on the GSA implementation by the Department of Statistics (<http://www-stat.stanford.edu/>). The chapter details the GSA Analysis, the algorithms to compute S-Scores and methods to explore the results of GSA analysis in **GeneSpring GX**.

24.2 Gene sets

A gene set from the Broad Institute is a group of genes, based on prior biological knowledge, that share a common biological function, chromosomal location or regulation. In **GeneSpring GX**, gene sets can also be defined as any entity lists created in the application that are used for GSA.

The Broad Institute (<http://www.broad.mit.edu/index.html>) maintains a collection of gene sets. **GeneSpring GX** supports the import of MIT-Harvard-Broad gene sets in the following file formats:

- **txt/csv:** First line is header information and the remaining lines are genes.

- **grp**: Gene set file format where each gene is in a new line
- **gmt**: Gene Matrix Transposed file format where each row represents a gene set
- **xml**: Molecular signature database file format (msigdb_*.xml)

A detailed description of the file formats can be found at http://www.broad.mit.edu/cancer/software/gsa/wiki/index.php/Data_formats. The Broad gene sets can be found at http://www.broad.mit.edu/gsa/msigdb/msigdb_index.html. Each individual gene set can be viewed, downloaded and imported into **GeneSpring GX**. Alternatively, after registering with the web-site, one can download the entire collection. Links to other gene sets can be found at (<http://www-stat.stanford.edu/~tibs/GSA/>).

Once Broad gene sets have been downloaded, they can be imported into **GeneSpring GX**. To import the Broad gene sets, click on the *Import BROAD GSEA Gene sets* link from the **Tools** section of the menu bar.

Importing gene sets in .grp, .gmt or .xml formats into **GeneSpring GX** converts them into **GeneSpring GX** Gene Lists which are automatically marked as Gene Symbol. (Note that importing the msigdb_v2.xml into **GeneSpring GX** takes around 10 minutes as the XML file is parsed)

Note: To perform GSA, the Entrez ID or Gene Symbol mark is essential. These are derived from the technology of the experiment. For Affymetrix, Agilent and Illumina technologies, **GeneSpring GX** packages the Entrez ID and Gene Symbol IDs marks. For custom technologies, Entrez ID or Gene Symbol must be imported and marked while creating custom technology for using the GSA.

24.3 Performing GSA in GeneSpring GX

GSA can be accessed from most of the workflows in **GeneSpring GX**. Clicking on the *GSA* link in the *Result Interpretations* section of the Workflow panel will launch a wizard that will guide you through GSA in **GeneSpring GX**.

Input Parameters The input parameters for GSA analysis is an entity list and an interpretation in the current active experiment. By default, the active entity list and the active interpretation in the experiment are selected. Clicking on the *Choose* option will show a tree of entity lists or interpretations in the experiment. You can choose any of the entity lists and interpretation from the tree as inputs to the GSA Analysis. See figure 24.1.

Pairing Options In the Pairing Options page, you can explicitly select pairs of conditions for GSA, or, you can select all the conditions in the interpretation against a single control condition. If you choose pairs of conditions, the table shows all the pairs. Choose the pairs of conditions to test by checking

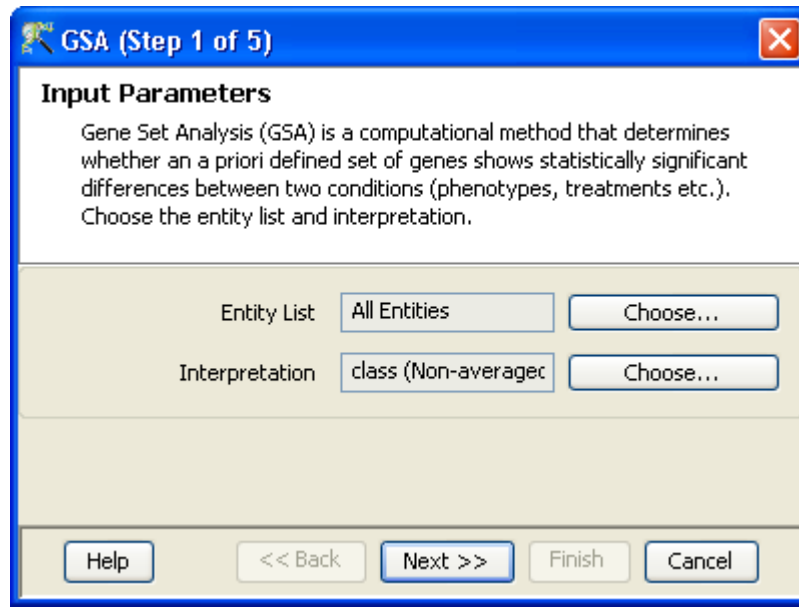


Figure 24.1: Input Parameters

off the corresponding boxes. If you choose all conditions against control, select the condition to use as the control from the drop-down menu. See figure 24.2.

Choose Gene Sets In the *Choose Gene Sets* options page, you can choose one or more of the BROAD gene sets that have been imported. See figure 24.3. Alternatively, you can select custom gene sets from entity lists that you have created in **GeneSpring GX**. To do this, click on the Advanced Search radio button, search for the entity lists of interest, and select the ones to be used as gene sets for GSA. See figure 24.4.

You can also specify the minimum number of genes that must match between the gene set and the input entity list for GSA in order for the gene set to be considered in the analysis. The default is set at 15 genes. Thus, if a gene set has less than 15 genes matching the entity list, then this gene set will not be considered. The default number of permutations used for analysis is set at 100.

Results from GSA The *Gene Sets satisfying minimum Gene requirement* spreadsheet shows the gene sets with p-values below the specified cut-off. The *Gene Sets with fewer than minimum necessary matches* spreadsheet shows the gene sets with p-values above the specified cut-off. You can change the p-value cut-off by clicking on the Change p-value cut-off button and entering a new value. See figure 24.5.

GSA results spreadsheet reports the following columns of values:

- **Gene Sets:** List of gene sets that pass the threshold criterion.
- **Details:** User supplied description associated with the gene set.
- **Total Genes:** Total number of genes in the gene set.

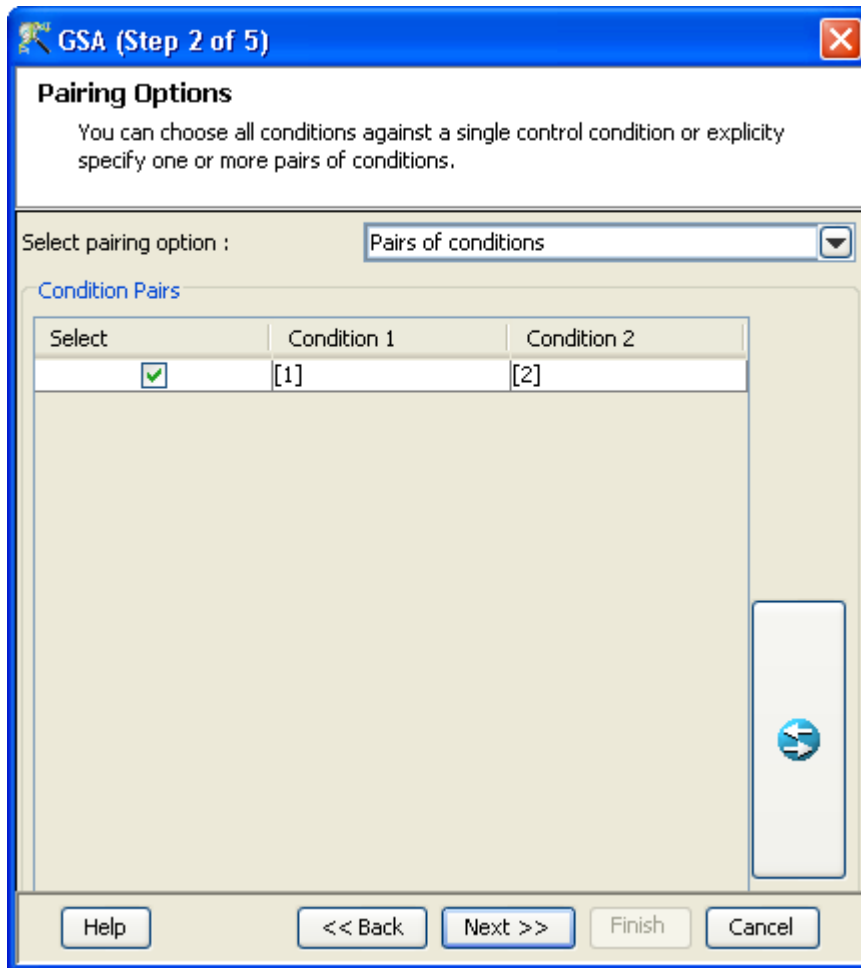


Figure 24.2: Pairing Options

- **Genes Found:** Number of gene in the gene set that are also present in the dataset on which analysis is performed.
- **S-Score:** Enrichment score of the gene set for the indicated pairs of conditions.
- **p-value:** Nominal p-value (from null-distribution of the gene-set)
- **Corrected p-value:** False Discovery Rate p-value, if p-value correction is requested

Columns except the first four are repeated when multiple pairs of conditions are selected for analysis.

Gene sets with Corrected p-values below the cut-off can be saved to the Navigator. Click *Finish* to save all the gene sets within the *Gene Sets satisfying minimum Gene requirement* spreadsheet. To save a subset of these gene sets, select the gene sets of interest and click Save Custom Lists. These gene sets will be automatically translated to the technology of the experiment and saved as entity lists in a GSA folder within the Navigator. The saved entity lists are named according their respective gene set names.

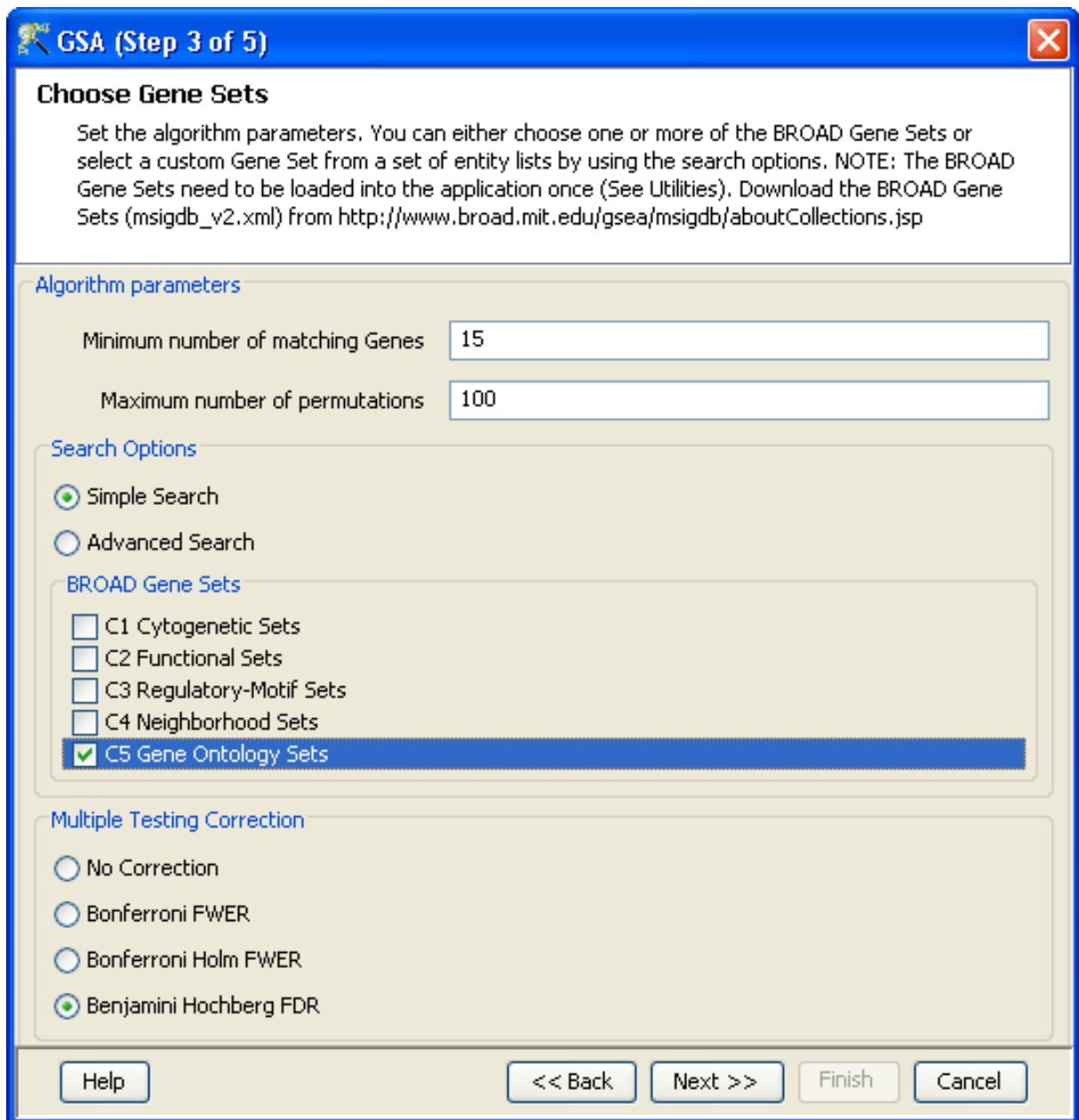


Figure 24.3: Choose Gene Sets

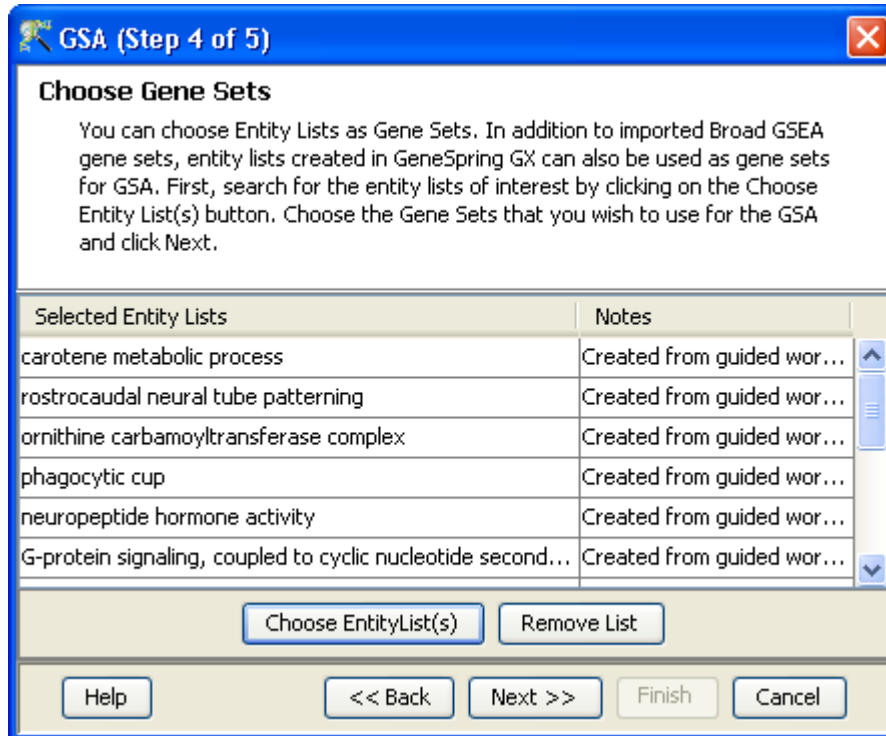


Figure 24.4: Choose Gene Sets

24.4 GSA Computation

Algorithm and computation of associated metric is detailed in the paper <http://www-stat.stanford.edu/~tibs/ftp/GSA.pdf>.

GSA analysis works on the t-Score computed for a list of genes, and uses a sum of t-Scores to compute the s-Score for each for gene set. t-Score in **GeneSpring GX** is computed for each gene assuming unequal variance between the conditions. The max-mean statistic is used to compute S-Scores for each gene set. The permutative procedure described in the paper is used to compute the *p-values*. S-Score for a gene set is compared with scores generated from random permutations of samples between the phenotypes to determine the level of significance. Number of permutations can be configured at *Tools* → *Options* → *Data Analysis Algorithms* → *GSEA* of the menu bar.

Analysis is restricted to log summarized datasets. If a gene has multiple probes in the dataset, the probe with maximum inter-quartile range across the expression values (with no more than 25% of the values missing) is used to represent the gene. Inter-quartile range is immune to baseline transformation and hence GSA results on data with and without baseline transformation remain the same.

Pseudo code for GSA in **GeneSpring GX**
 $R = \text{NumberOfRowsInDataset}$

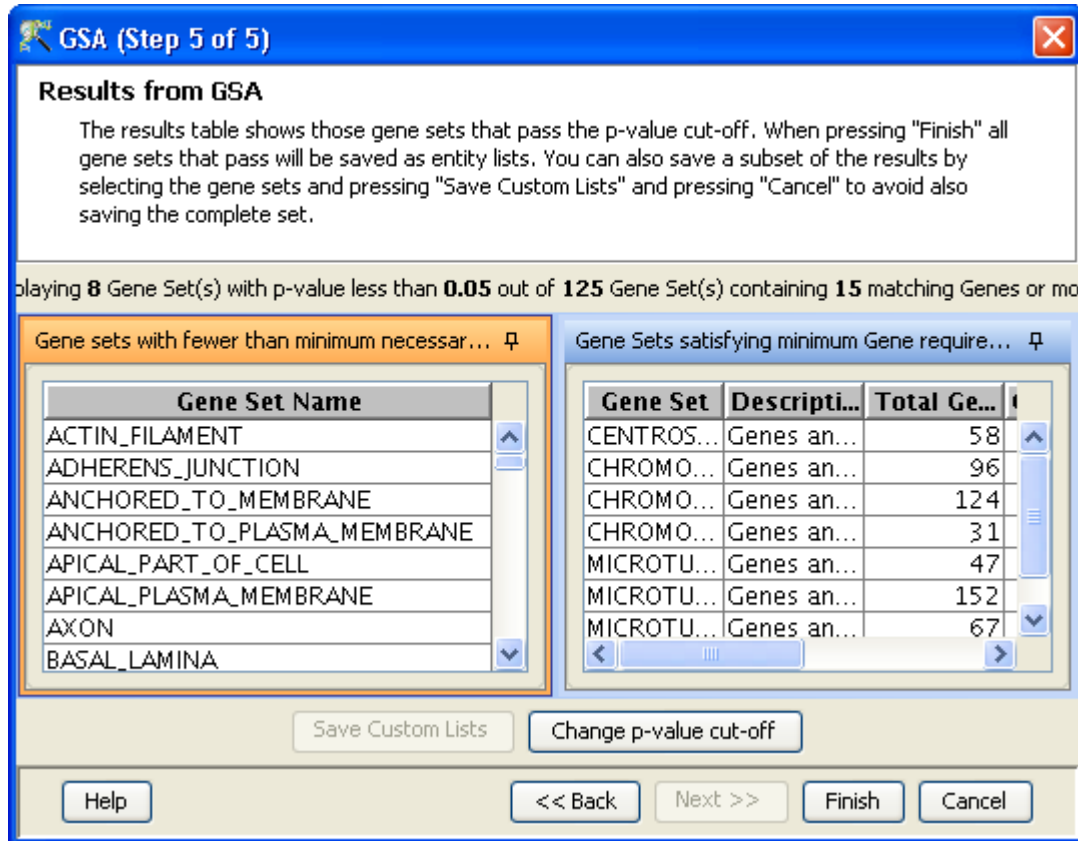


Figure 24.5: Choose Gene Lists

Create a collapsed with rows with unique genes for a given pair of phenotypes

For each row in the dataset

$$n_1 = \text{NumberOfSamplesOfPhenotype}_1$$

$$n_2 = \text{NumberOfSamplesOfPhenotype}_2$$

$$m_1 = \text{mean}(\text{phenotype}_1)$$

$$m_2 = \text{mean}(\text{phenotype}_2)$$

$$\text{var}_1 = \text{variance}(\text{phenotype}_1)$$

$$\text{var}_2 = \text{variance}(\text{phenotype}_2)$$

$$t\text{Score} = \frac{m_1 - m_2}{\sqrt{\frac{\text{var}_1}{n_1} + \frac{\text{var}_2}{n_2}}}$$

Separate the positive and negative t-scores into two arrays

$$t(> 0) = \text{ArrayOfPositivetScores}$$

$$t(< 0) = \text{ArrayOfNegativetScores}$$

$$\text{mean}_+ = \frac{\text{sum}(t(>0))}{R}$$

$$\text{mean}_- = \frac{\text{sum}(-t(<0))}{R}$$

$$\text{stdev}_+ = \sqrt{\frac{\text{sum}(t(>0)^2) - R * \text{mean}_+^2}{R}}$$

$$\text{stdev}_- = \sqrt{\frac{\text{sum}(t(<0)^2) - R * \text{mean}_-^2}{R}}$$

For each gene set

$g = \text{NumberOfGenesInGeneSet}$

$$sScore_+ = \frac{\frac{\text{sum}(t(>0)\text{ForGenesInGeneSet}) - \text{mean}_+}{g}}{\text{stdev}_+}$$

$$sScore_- = \frac{\frac{\text{sum}(-t(<0)\text{ForGenesInGeneSet}) - \text{mean}_-}{g}}{\text{stdev}_-}$$

$sScore\text{ForGeneSet} = \max(sScore_+, sScore_-)$ Compute $sScore_*$ for each random permutation of phenotype ordering

$$p - \text{value} = \frac{\text{NumberOfPermutationsWhere}(\text{abs}(sScore) < \text{abs}(sScore_*))}{\text{TotalNumberOfPermutations}}$$

Chapter 25

Pathway Analysis

25.1 Introduction to Pathway Analysis

Traditional analysis of gene expression microarray data involves applying statistical analysis to identify genes that are differentially expressed between the experimental conditions. However, it is difficult to extract a unifying biological theme from a list of individual genes that is obtained from such statistical analysis. Thus, after identifying genes of interest in **GeneSpring GX**, it is often desirable to put these statistically significant findings into a biological context.

GeneSpring GX now supports pathway analysis via the following:

- A database of biological and chemical entities, relationships between entities, and properties of these entities and relationships.
- A set of pathway creation algorithms which query this database or other sources of literature with a specified list of entities.
- An interactive pathway viewer which allows visualization of these pathways and overlay of data on these pathways.
- A *Find Significant Pathway* function that helps determine which of the created pathways have significant overlap with a specified list of entities.
- Given a 'Term', pull out all interactions containing the MeSH terms associated with this term and create pathways based on those interactions. Alternate way of creating pathways based on terms/concepts instead of entities.

GeneSpring GX can automatically map the entities within a user selected Entity List or Gene list to the genes in the Pathway database. This allows user to integrate information regarding the dynamics and

dependencies of the genes within a pathway and how their expressions change across your experimental conditions. The Pathways tool allows you to quickly answer the questions; What probable pathways are represented by my genes of interest? What is the hierarchy of function of my genes of interest? In which biological pathways is there a significant enrichment of my genes of interest? In doing so, you can quickly determine how the experimental conditions affect certain biological pathways and processes, and not just the expression of individual genes.

25.2 Licensing

The pathway features in **GeneSpring GX** are licensed separately as an additional module license. These features will work only if you have this additional module license.

If you have a desktop license for **GeneSpring GX** that does not include pathway features, contact **GeneSpring GX** support to get a pathway module license order-id. Then use *Help* → *License Manager* → *Change* and provide this order-id in the box specified. All pathway features will be available after authentication with the **GeneSpring GX** license server; **GeneSpring GX** will need to be restarted though.

If you have a Workgroup license that does not include pathway features, contact your systems administrator and request them to contact **GeneSpring GX** support to get a pathway module license order-id; this order-id can be used to upgrade the **GeneSpring GX** floating license server to include the pathway module. Once this is done, the next time you launch the **GeneSpring GX** client, all pathway functionality should be enabled.

25.3 Getting Started

GeneSpring GX supports several organism specific databases. Databases for Human, Mouse, Rat, Drosophila, Arabidopsis, C. elegans, E. coli, and Yeast are supported via updateable data packages. Go to *Annotations* → *Update Pathway Interactions* to update the database for the organism(s) of your choice.

Please ensure that you have enough disk space in your **GeneSpring GX** installation folder (a minimum of 10GB of disk space is needed to fit in all the above organisms) before you perform the above updates. If you do not have enough disk space in the **GeneSpring GX** installation folder then edit the following file `installdir/app/MySQL/my.ini` and modify the `datadir` parameter to provide a folder name which has sufficient space; then restart **GeneSpring GX** and perform the above updates.

Once the updates have been performed, clicking on *Annotations* → *Pathway Database Statistics* will confirm that the pathway databases of interest are indeed available, i.e., listed in the left panel, with associated statistics on the right. If the above step gives an error message, check that the MySQL database is running using the task monitor (in the user processes tab you should see `mysqld-nt`). Then contact **GeneSpring**

GeneSpring GX support with this information as well as the stderr.log and stdout.log files in the installation directory, and all files from the bin/launcher/lib/logs subfolder.

Note that in Workgroup mode, the *Annotations*→*Update Pathway Interactions* menu item is grayed out. Contact your Systems Administrator to update the Pathway Interactions Server with the appropriate organism databases.

25.4 Working with Other Organisms

If you wish to work with organisms other than Human, Mouse, Rat, Drosophila, Arabidopsis, C. Elegans, Yeast and E. Coli, then you will need to create a new organism via *Annotations*→*Create Pathway Organism*. You will need to provide the exact scientific name, common name and taxonomy identifier. Though **GeneSpring GX** does not perform correctness checks on these terms, it is useful to provide exact and valid names for smooth functioning later; for instance, use 9606 as taxonomy identifier, “Homo sapiens” (note the space between the two words and the small letter which begins the second word) as scientific name, and “Human” as common name. Valid taxonomy ids can be found at the NCBI Taxonomy site (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy>).

25.5 Pathway Analysis in Microarray Experiment

GeneSpring GX allows the user to perform pathway analysis on lists of entities selected from microarray analysis. The pathway analysis functionality can be launched in two ways:

- Within the microarray analysis workflow, which will be described in this section;
- Separately as a Pathway Experiment that is described in [Pathway Experiment](#) of this manual.

Pathway analysis can be launched within the microarray analysis workflow from *Results Interpretations*→*Pathway Analysis*. Any gene list derived from significance and fold-change analysis in the microarray experiment can be selected from the explorer pane and further studied in the context of pathways.

Organism specific relation databases are available as updates to **GeneSpring GX**. The relations in the database are mainly derived from published literature abstracts using a proprietary Natural Language Processing (NLP) algorithm. Additional interactions from experimental data, available in public repositories like IntAct are also included in the **GeneSpring GX** databases. Details about the interactions database are available in section [Pathway Database](#).

A few terms related to the **GeneSpring GX** pathway database are explained below:

Relation score: This property indicates a confidence matrix on the quality of relations in the pathway database in **GeneSpring GX**. All relations derived from curated databases like IntAct and all user created relations are given a score of 10 (highest score). For the NLP derived relation, the relations are graded on a scale of 1-9, the best being 9 and the weakest is 1. The score properties are internally calculated based on the number of references and the syntax of the sentences. NLP derived relations have lower score than any curated or user derived relations.

Each reference associated with an relation is graded on a scale of 1-9, based on the semantics rules of the NLP, the weakest being 1 and the strongest being 9. This property is called the *RefScore* of the reference. Any relation supported by at least one reference of Refscore 9 or having 3 or more references supporting it, is graded as 9.

Connectivity: Connectivity of an entity represents how well the entity is connected to other entities. In the context of a pathway view (network) the “local connectivity” of an entity is defined as the number of other entities in the view that are connected to it. “Global connectivity” of an entity is independent of the view and is defined as the total number of relations in which the entity participates.

While expanding a network, the users are given the option to limit the number of entities to be included in a pathway view. The user can either choose to expand based upon “local” connectivity or the ratio of its “local/global connectivity”.

25.5.1 Pathways, Entities and Relationships

Entities, Relationships and their properties reside in the pathway databases described above. A *pathway* on the other hand denotes a collection of entities and associated relationships. Pathways are created by querying the relations databases and are saved inside **GeneSpring GX**; they can then be added to individual experiment navigators.

Note that each pathway has an associated organism and a pathway can be added to the navigator of an experiment only if the organism of the experiment matches the organism of the pathway. The organism of a pathway can be determined by inspecting that pathway and the organism for an experiment can be determined by inspecting its corresponding technology. Note that the former is non-editable while the latter is indeed editable and can be modified in rare instances where there is a mismatch.

25.5.2 Analysis

The user can query the relations database with a specified entity list and create a pathway with entities in the starting entity list and other related entities in the database. A variety of algorithms are available to do this. In each case, the database that is queried corresponds to the organism of the technology of the current experiment. The Entrez ID, SwissProt ID and Gene Symbol from the technology are used for this query and hence it is important that both the technology and the relations database contain at least one of these properties.

GeneSpring GX allows the user two different options for pathway analysis:

1. *Simple or Guided Analysis*: This option aims to allow the user to explore the most common functionalities of a pathway analysis. The default settings for guiding the user through a simple pathway analysis include:
 - matching the entity list of interest to entities in the database
 - retrieving relevant relations between the set of matched entities
 - displaying the results in a graphical view in the form of a network.
2. *Advanced Analysis*: This option aims to allow the user to explore all the functionalities of a pathway analysis in detail and change the settings at every step of the analysis, as required.

The analysis steps are listed below:

1. **Step 1 of 5:**

The first step allows the user to select an entity list of interest and enter the parameters for analysis. The wizard termed **Input Parameters** allows the user to choose the analysis and algorithm for building the pathway. For the guided pathway analysis, the user has to select 'Simple' as the type of analysis. For changing the parameters for pathway creation, the user has to select 'Advanced' type of analysis. The user has to choose the network algorithm that will be used to query the relations database and create a pathway network from the selected entities.

The available options for *Simple Analysis* workflow are:

- **Direct Interactions**: Finds relations that connect the entities in the selected entity list.
- **Network Targets and Regulators**: Finds entities that are upstream and downstream of two or more entities from the original list.
- **Network Targets**: Finds downstream entity targets that connect two or more entities from the original list.
- **Network Regulators**: Finds upstream entity regulators that connect to two or more entities from the original list.
- **Network Binders**: Finds entities that 'bind' (connected by binding interactions) to two or more entities from the original entity list.
- **Network Modifiers**: Finds protein entities that are either regulators or targets of biochemical protein modifications (eg. Phosphorylation, ubiquitination, etc.) of two or more proteins from the original entity list.
- **Transcription Regulators**: Find protein entities regulating mRNA expression of, or whose expressions are regulated by, two or more entities from the original list.
- **Transport Regulators**: Finds all molecules that are regulating the transport of other molecules.
- **Metabolism Regulators**: Molecules that are regulating the metabolism of biomolecules.
- **Small Molecules**: Finds all small molecules (drugs) regulators and targets of two or more entities from the original list.

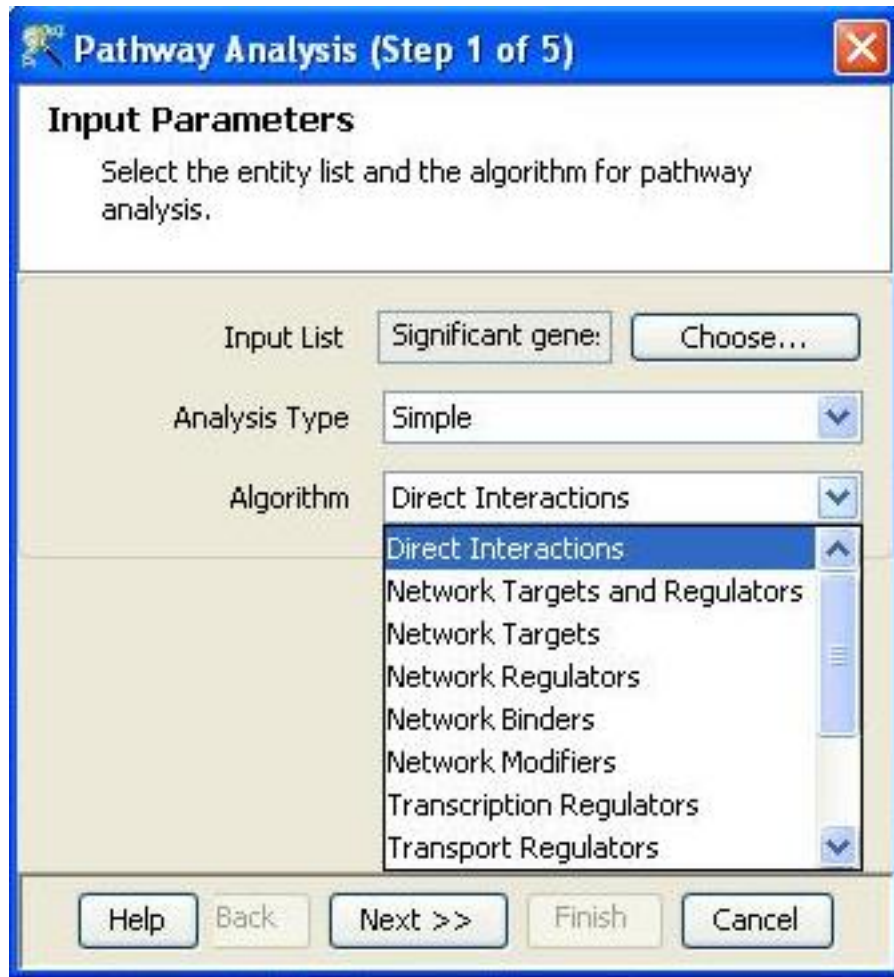


Figure 25.1: Simple Analysis

- **Biological Processes:** Finds all biological process entities connected to two or more entities from the original list.
- **Shortest Connect:** Finds the smallest set of relations that will connect all entities in a given list into a single network.

See figure 25.1.

For the *Advanced* type of analysis, the user has to select between the three algorithms:

- **Direct:** Finds relations that connect the entities in the selected entity list.
- **Expand:** Expands the existing network to include the first-degree neighbors of the selected entities.
- **Shortest Connect:** Finds the smallest set of relations that would connect a set of entities into a single network. Note that some intermediate entities may be introduced in this process.

See figure 25.2

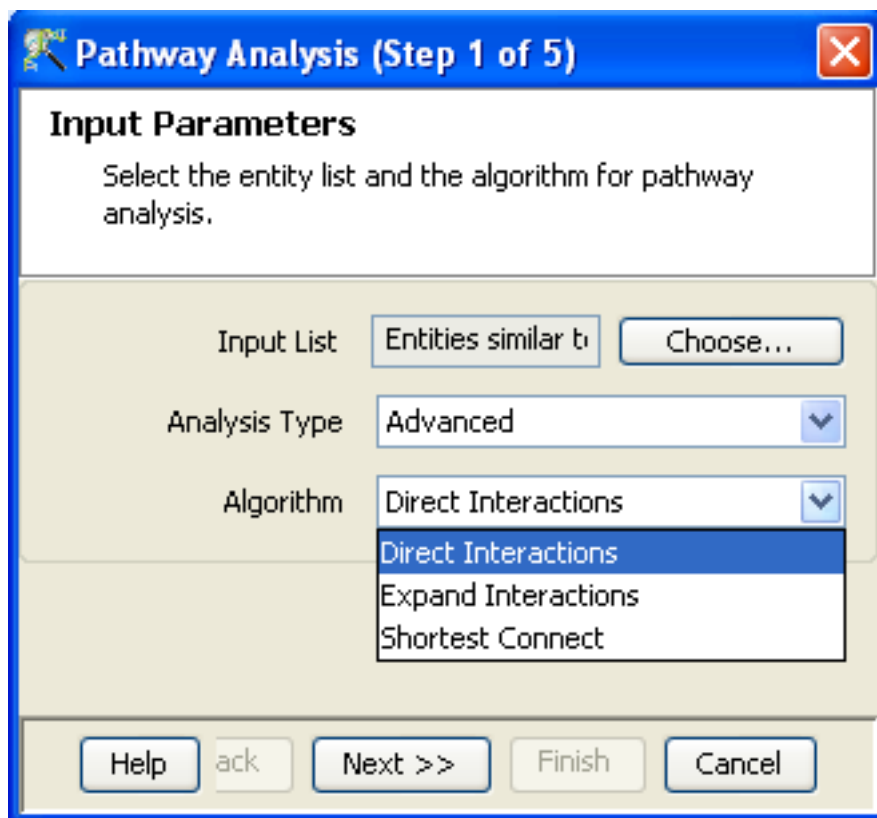


Figure 25.2: Advanced Analysis

If the selected algorithm does not find any entity that honors the criteria, an error message appears on the screen indicating this and the user can select a different algorithm or another entity list for analysis.

See figure 25.3

2. Step 2 of 5:

This wizard displays the **Matching Statistics** after matching entities in the input gene list to entities in the database. The “Match Result” column indicates whether a match was found or not. For matched rows, additional details like Probe Set Id, Name (gene symbol), Type (of matched entity), and Global Connectivity (total number of relations in which the matched entity participates) are also displayed.

This wizard step is displayed only in the advanced analysis. A subset of the matched rows can be selected before moving to the next step.

See figure 25.4

3. Step 3 of 5:

This wizard termed the **Analysis Filters** appears only in case of *Advanced Analysis*. This allows the user to set filters on the following parameters:

For **Direct Algorithm** in the *Advanced Analysis* workflow, the following filter settings are allowed to create a network:

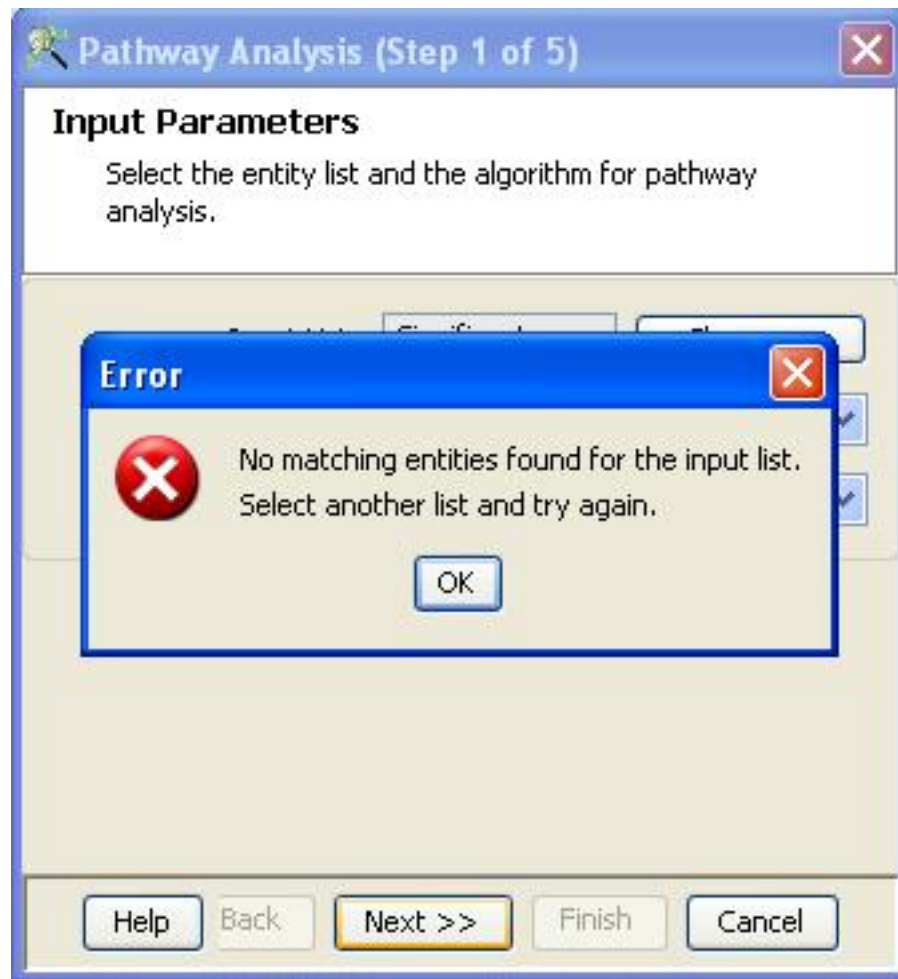


Figure 25.3: Error Message

Relation Filter:

- *Relation score*: The user can build or expand a network by including only relations with a particular quality score. The concept of assigning a quality score to every relation is explained above. The default score defining relation quality is set to ≥ 9 . See section on score above for details.
- *Relation types*: The user can select one or more types of relations (see relation types below for details on each type) to generate the network from the starting list of entities, depending upon the biology of the problem. Check boxes are provided for multiple selections. The default settings use all type of relations to create the network. Eg. To create a network of transcriptional analysis, the user can select only 'expression' and 'promoter binding' types of relations.

See figure 25.5.

For **Expand Algorithm** in the *Advanced Analysis* workflow, the following filter settings are allowed to create a network:

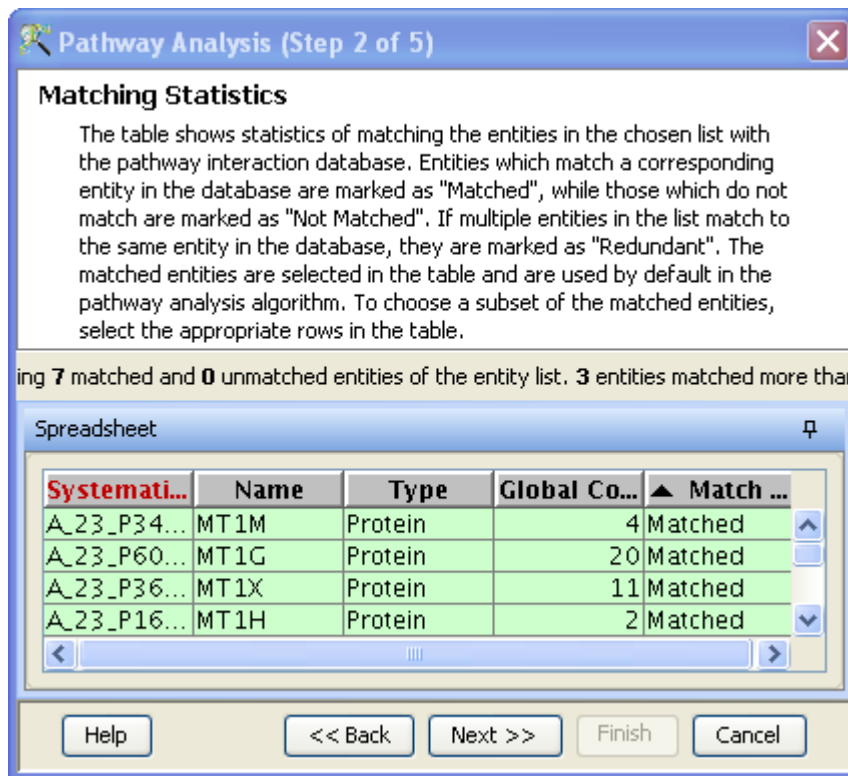


Figure 25.4: Matching Statistics

- **Relation Filter:** The usability of this filter is described above in the *Direct Algorithm* in the *Advanced Analysis*.
- **Entity Filter:**
 - *Entity local connectivity:* This filter allows the user to add new entities to a given network by ranking the new entities base upon their local connectivity. The default settings are > 2 , to consider only those entities which are connected to 2 or more entities from the starting list.
 - *Entity type:* The user can preferentially select the types of entities to add to the expanded network, depending upon the biological question. Check boxes are provided for multiple selections. The default settings use all type of entities to create the network. Eg: The user can opt to study all biological processes and functions regulating the list of genes. In this case, the user will expand the list by limiting the newly added entities to Processes and Functions.
- **Limit analysis results based on**
 - *Local connectivity:* Allows user to add a certain number entities to the given network based upon their rank on local connectivity. New entities are ranked with decreasing priority, based upon how many entities they connect within a given list of entities.
 - *Local to global connectivity ratio:* A local/global connectivity ratio is computed for each new entity. Local connectivity is based upon the number of entities to which it connects within a given list and global connectivity is the number of relations that it participates

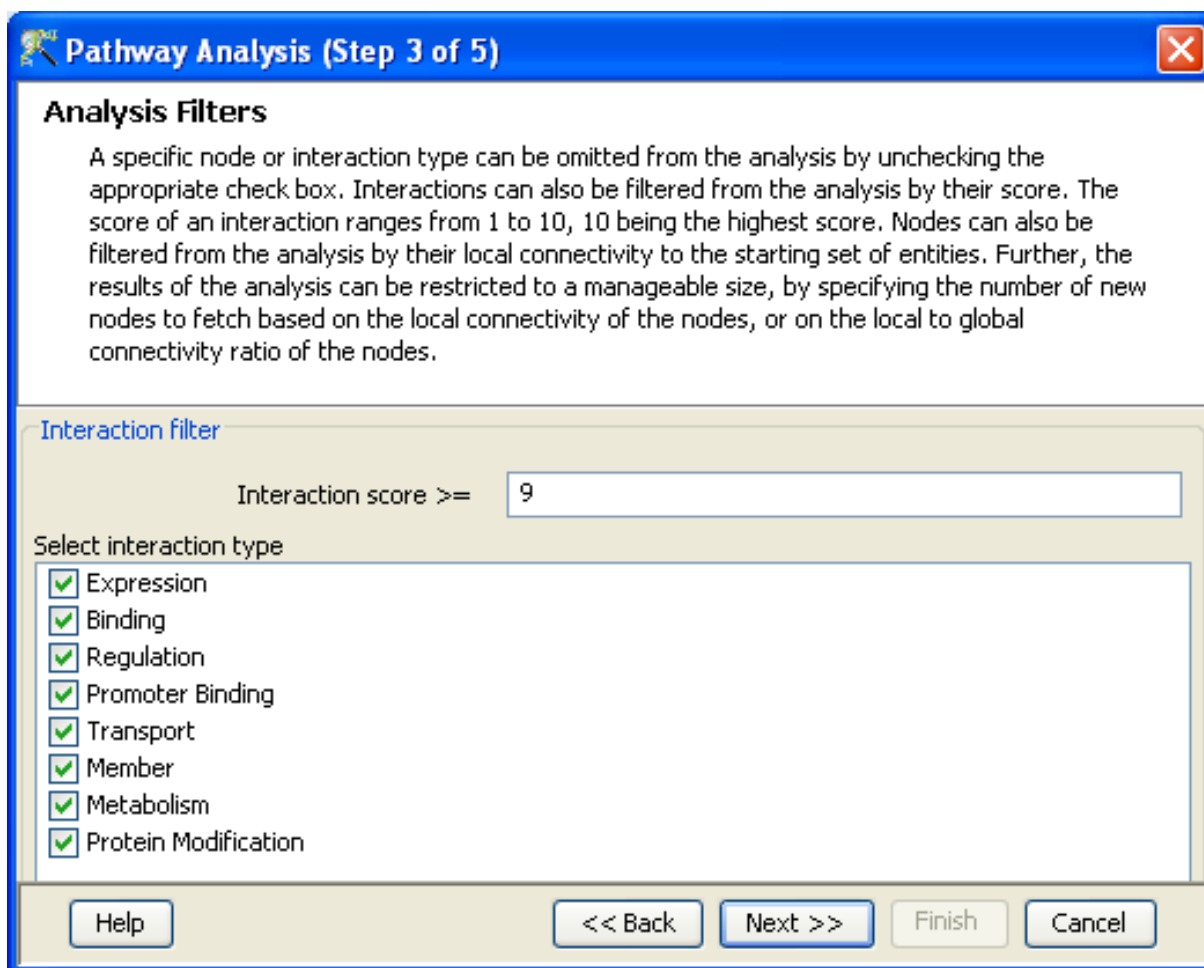


Figure 25.5: Analysis Filters-Direct Algorithm

in the entire database. New entities are ranked with decreasing priority, based upon this local/ global connectivity ratio.

The default is set to Local to global connectivity ratio

- **Maximum number of new entities:** This allows the user to limit the maximum number of entities to be added to a network. The default is set to 50.

See figure 25.6.

For **Shortest Connect Algorithm** in the *Advanced Analysis* workflow, the following filter settings are allowed to create a network:

- **Relation Filter:** The usability of this filter is described above in the *Direct Algorithm* in the *Advanced Analysis*.
- **Entity Filter:**
 - Entity global connectivity:** This filter allows the user to add new entities to connect two disconnected network clusters by ranking the new entities base upon their global connectivity. The default settings are 100 and above.

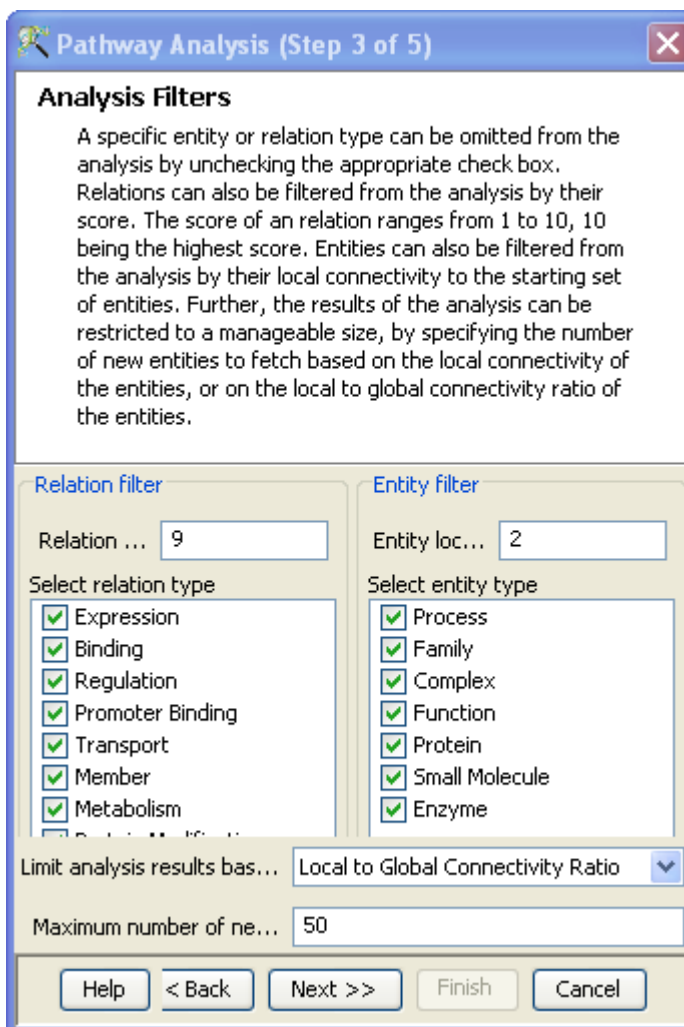


Figure 25.6: Analysis Filters-Expand Algorithm

Entity type: The filters are set to the same as described in the **Expand Algorithm** of Advanced analysis.

Tip: Internally the algorithm tries to create a single connected component by carrying out a series of expansions. If the number of expansion steps is more than 10 or if the number of entities from which expansion needs to be carried out next exceeds 10,000, the algorithm will abort. Some small molecules, such as calcium, have very high connectivities (greater than 5000). If these molecules are encountered in the initial stages, the algorithm will probably abort because of the large size of the expansion set. Two common tips to make the algorithm find the shortest path:

- Use the filter to turn off the the Small molecules Entity type. This prevents Calcium and Glucose from jamming shortest path.
- Increase the global connectivity of your protein entities. the default is set to 100. You could increase this to 1000 and simultaneously turn off the small molecules.

See figure 25.7

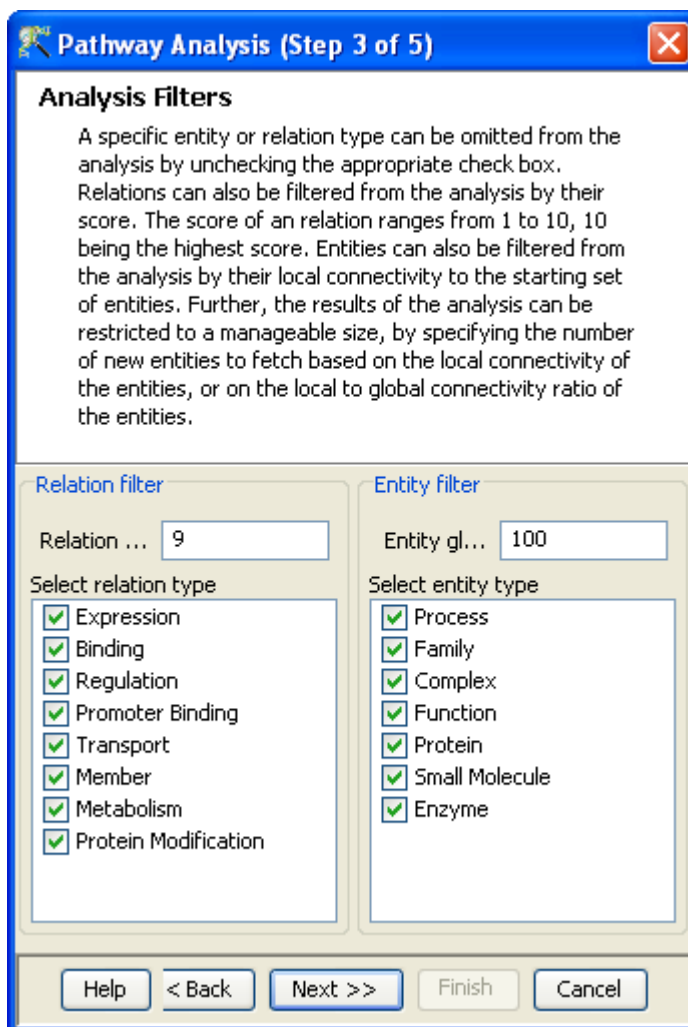


Figure 25.7: Analysis Filters-Shortest Connect

4. Step 4 of 5:

The view is called **Analysis Result**. This wizard view displays the created pathway. The initial number of entities, the number of new relations and the number of new entities are displayed.

See figure [25.8](#)

5. Step 5 of 5:

This **Save Pathway** wizard summarizes the pathway created. Here the user can specify a name, and add to the notes that are generated listing the steps used in creating the pathways. The user can click *Finish* to create the Pathway view saved in the explorer pane in a branch in the tree, under the selected list.

See figure [25.9](#)

Clicking on the corresponding icon in the explorer can launch the pathway view.

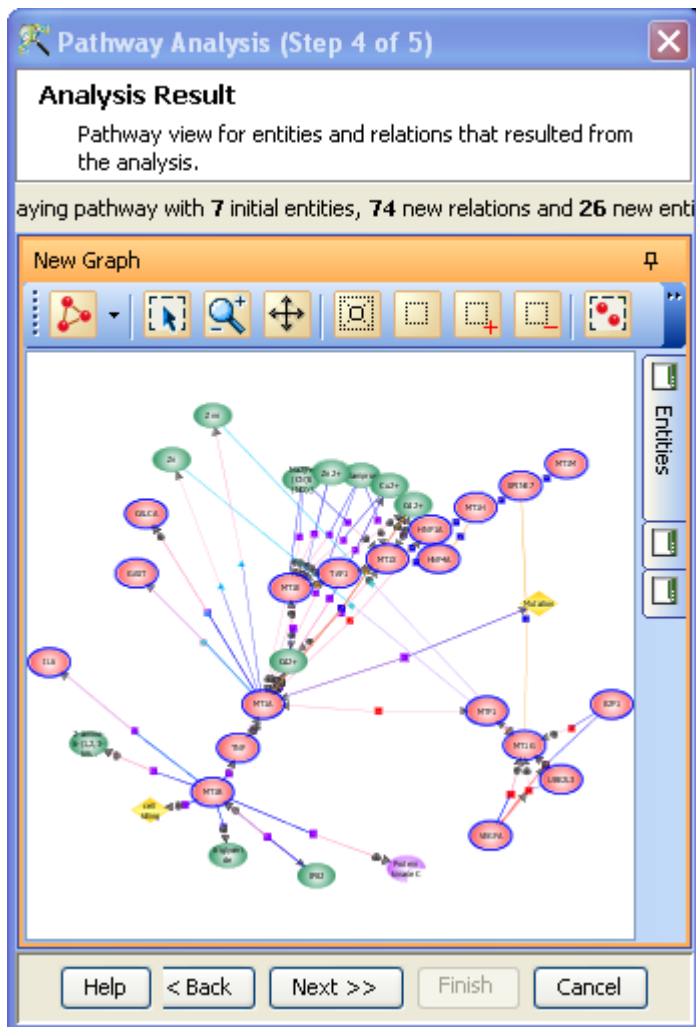


Figure 25.8: Analysis Result

The Simple Analysis launches only the Wizard for Step 1 of 5. After the user has specified the algorithm, the filters for Simple analysis are set to default:

- **Algorithms type:** local/ global.
- **Connectivity :** Connectivity relevance: 50 Connectivity: ≤ 2
- **Entity filter:** All entities are selected.
- **Quality filter:** ≥ 9
- **Relation filter:** All relation types are selected.

The *Simple Analysis* proceeds directly from step 1 to step 4 and then launches the pathway.

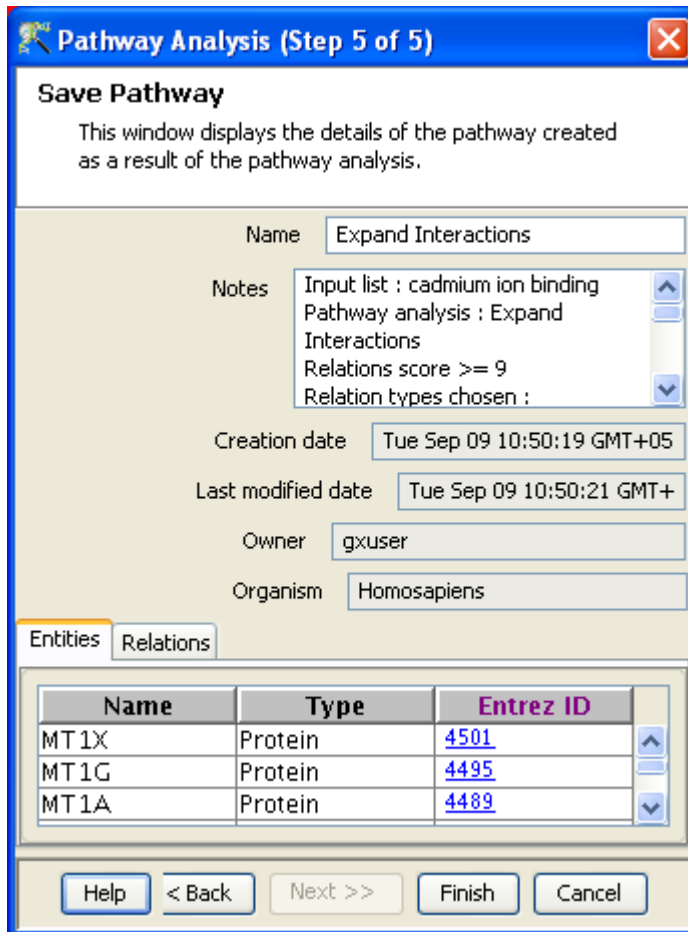


Figure 25.9: Save Pathway

25.5.3 Pathway View

A simple pathway view appears with the network (graphical) representation of the biological pathways. Nodes of the graph could be any of the following. The node legend displays the default thematic settings for each type of entity. See Figure 25.10.

- PROTEINS/ GENES
- SMALL MOLECULES
- ENZYMES
- PROCESSES
- FUNCTIONS
- COMPLEX

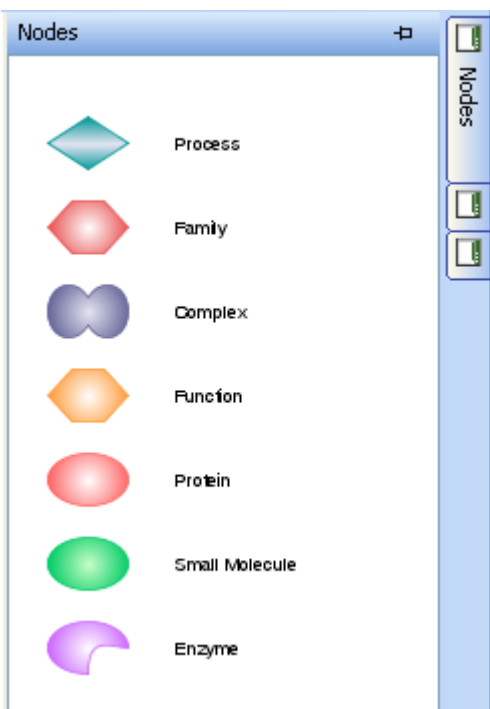


Figure 25.10: Node-Legend

- **FAMILY**

The user can review the information of each entity by launching its property table from the UI. Double clicking on a node will open its property table. See figure 25.11. For details on the properties of each entity, refer to section [Database Entities](#).

Relations between various molecules and/or processes are represented by edges of the network. See Figures 25.12 and 25.13. The molecular relations that are distinguished in **GeneSpring GX** are the following:

- **BINDING**
- **REGULATION**
- **PROTEIN MODIFICATION**
- **EXPRESSION**
- **PROMOTER BINDING**
- **METABOLISM**
- **TRANSPORT**

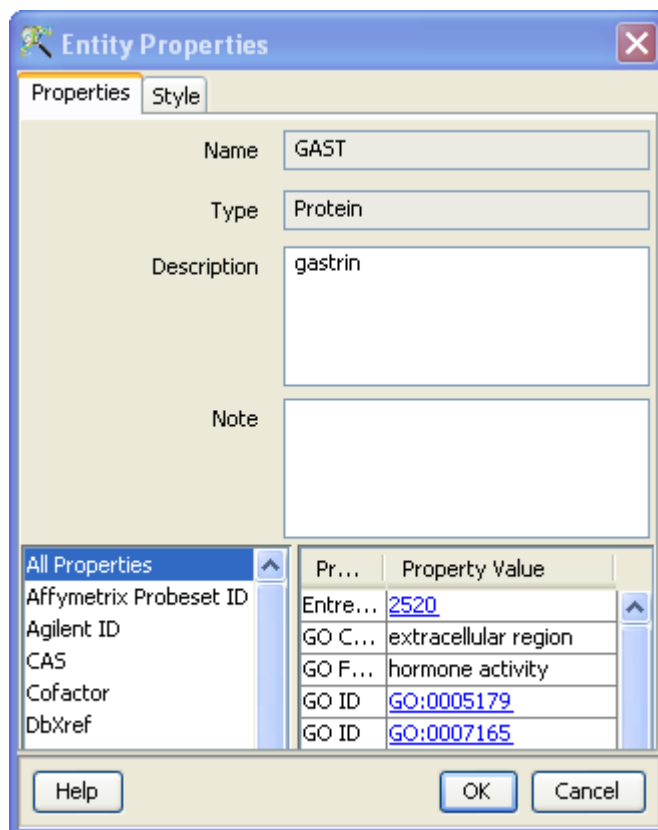


Figure 25.11: Node Properties

- **MEMBER**

Each relation has several properties that can be viewed in a table by double clicking on the relation. See figure 25.14. Details of the Relation properties are discussed in Section [Relations](#).

Briefly, each relation is characterized by the following:

1. Participating entities and their roles (see Table on [Relations](#) for more details). The edges of the relation specify the roles of the participating entities. The color and representation settings can be altered by the user by using the **Theme** option.
2. Mechanisms (eg. phosphorylation, ubiquitination etc.)
3. Score that determines the quality of the relation. Number of references, source of the data and a reference score for each associated reference all contribute to a relation score.
4. Each reference that contributes to the relation is associated with the Journal, Year of publication and a reference score. The exact sentence that was interpreted by the NLP can also be viewed.

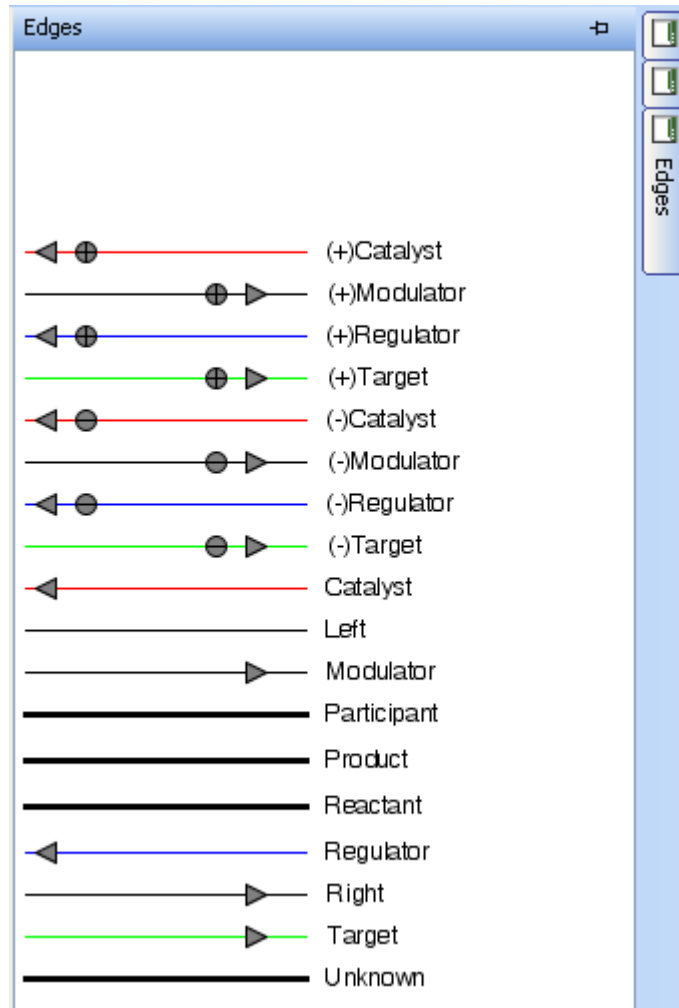


Figure 25.12: Edges-Legend

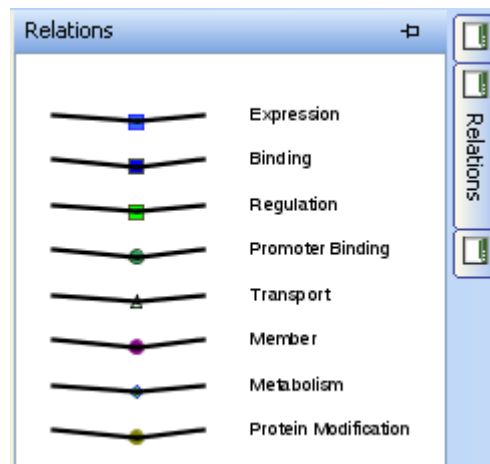


Figure 25.13: Relations-Legend

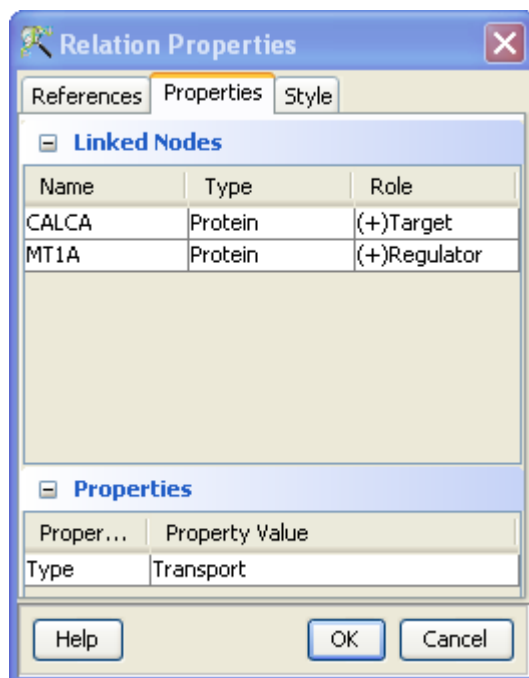



Figure 25.14: Relation Properties



Figure 25.15: Toolbar

Pathway Viewer

To launch a new pathway viewer, the user can click on the Create New Pathway  icon from the main tool bar. The user can create custom pathways or copy paste pathways onto this new view. In a pathway view, the entities circled in blue match the entities having Entrez IDs in the currently selected entity list in the project navigator. However, when matching objects from the database, other relevant properties of entities (for example-SwissProt ID) are also considered.

- **Toolbar:** The pathway viewer has a tool bar with several pathway building functionalities for the user. See figure 25.15.



Layout graph: Change the visual layout of the network.



Save pathway: Saves the pathway from the active Pathway viewer.



Selection mode: Select entities and relations.



Zoom mode: Magnify or reduce the image.



Pan mode: Move image while retaining all aspects of the image.



Select all: Selects all entities and relations in the viewer.



Remove unlinked entities: Unlinked entities are deleted from the Pathway viewer.



Zoom to fit view to visible area: Zooms the active pathway within the Pathway viewer.



Zoom selected region: Selects a portion of the Pathway viewer and zooms in.



Zoom in: Enlarges the Pathway view.



Zoom out: Reduces the Pathway view.



Copy selection: Copies all selected entities and relations to the clipboard.



Paste selection: Pastes copied selection to the active Pathway viewer.



Redo: Helps to redo an action



Undo: Helps to undo an action.

- **Right-click legend:**

Right click on the pathway view launches a legend with the following functionalities explained in the below table:

Drop down menu	Sub-menu	Details
Pathway Analysis		Launches the Analysis wizard. In this case the starting point for analysis is the set of selected entities in the view.
Magnifier		Magnifies to focus on a subsection of the graph
Show report	Entities table	Shows the Object ID, entity type, name, description and connectivity of all the nodes from the network in the form of a table.
	Selected Entities table	Shows the above properties of only the selected entities from the network in the form of a table
	Relation table	Shows the IDs, relation types, effects and the participating entities for all the entities in the network in the form of a table
	Selected Relation table	Shows the above properties for only selected relations and entities in the form of a table
Filter by type		Filters on types of entities
Reset filter		Resets the default filter options
Search by property		Searches the current view based upon properties of the entities
Delete selection		Delete selected entities
Merge selection		Collapses selected entities into a single node with user defined name
Expand selection		Expands the network to include first degree neighbors of the initial entity set.
Export	Image	Allows export in file formats like .png, .tiff, .jpeg, .bmp etc.
	Navigable html	Saves an html file where each entity and relation is hyperlinked to its property
Properties		Opens the tab for setting parameters for overlaying associated data on the network.

Table 25.1: Right-Click Legend

- *Export as*

- **Image:** A wizard that allows alterations in image attributes such as size and resolution is shown. Users can save images in file formats like .png, .jpeg, .jpg etc. If the user chooses to export the full image, the resulting image will have the all entities and controls drawn with the sizes

specified by the theme in use. If the pathway has a large number of entities, then exporting the full image can be very memory intensive. In this case, the user can choose not to export the full image - the exported image is then exactly the same as what is visible in the pathway view.

- **Navigable HTML:** A view can also be exported as a navigable html page via right click on the viewer and selecting *Tools* → *Export As* → *Navigable HTML* from the drop-down menu. This will create an html file where each entity and relation is hyper-linked to its properties. In order to transfer the saved html file, the user will need to carry the properties folder along with the image file.

- **Data Overlay:**

One of the key utilities of **GeneSpring GX Pathway Analysis** module is to dynamically view the microarray data associated with the genes in the context of their biological pathway connection. The user can launch the data overlay functionality from drop down *Right-click* → *Properties* → *Parameters*. See figure 25.16. The user has to select -

- **Show List Overlay:** This option allows user to select the entity lists from whose associated data should be overlaid on the pathway.
- **Show Interpretation Overlay:** This option allows the pathway to display overlay (averaged or non-averaged) over any treatment.

A heatmap is generated for every node to depict the signal intensity of gene expression for each experimental group and every replicate within the group. See figure 25.17

The scale for the overlaid data appears in the legend view. The user can alter the scale and the choice of colors. See figure 25.18.

- **Search Entities and Pathways:**

In addition to querying the database for connections amongst entities of interest, the user can search specifically for entities and pathways. The user can search from the main menu drop down options: *Search* → *Entities(Pathways)*. See figure 25.19

- **Search** → **Entities**

1. **Step 1 of 3:** The first step prompts the user to search any Entity property field for terms of interest. See Figure 25.20
2. **Step 2 of 3:**
The search results from Step 1 are displayed in the Output views. See Figure 25.21
3. **Step 3 of 3:**

The search results are displayed on the Entity Inspector. See Figure 25.22

- **Search** → **Pathways**

A wizard prompts the user to perform either:

1. Simple Search or
2. Advanced Search

Given a search term, the simple search performs search in every property field of an entity or pathway. Advanced search allows the user to set search on specific property fields.

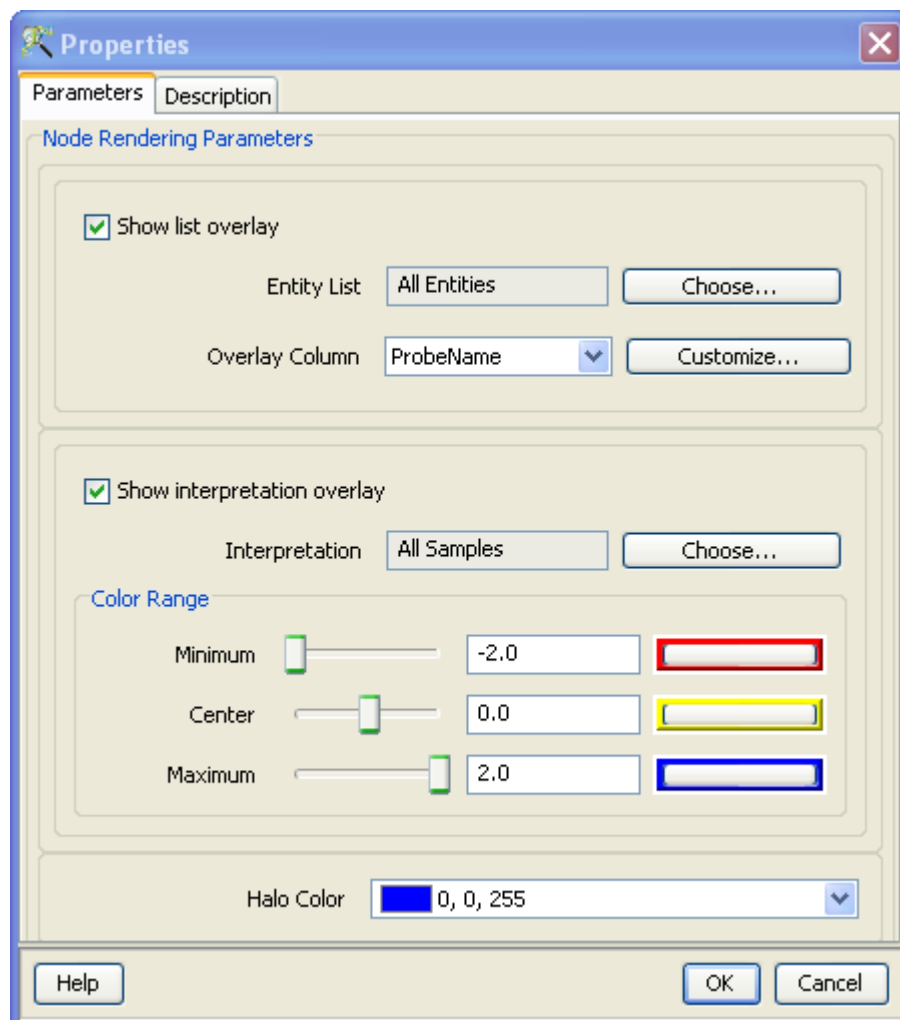


Figure 25.16: Data Overlay Properties

1. **Step 1 of 3:** The first step prompts the user to select either a Simple or an Advanced Search. See Figure 25.23
2. **Step 2 of 3:**
This view appears only in Advanced Search settings. The Advanced Search Parameters allows the user to search pathways based upon date of creation/modification, name, organism and attributes of participant entities. Several searches can be concatenated. See Figure 25.24
3. **Step 3 of 3:**
The search results are displayed on the Search results wizard. See Figure 25.25
A simple search proceeds directly from step 1 to step 3.

In addition, users can search an entity by dragging the appropriate symbol from the legend. Dragging any node will prompt the user to enter a search term. The pathway tool will try to search the existing database to find a matching entity and display all the matches in a table. The user can choose the entities from this table, which will be automatically displayed on the viewer. If the given search does not find a match in the database, it will create a new entity for the view.

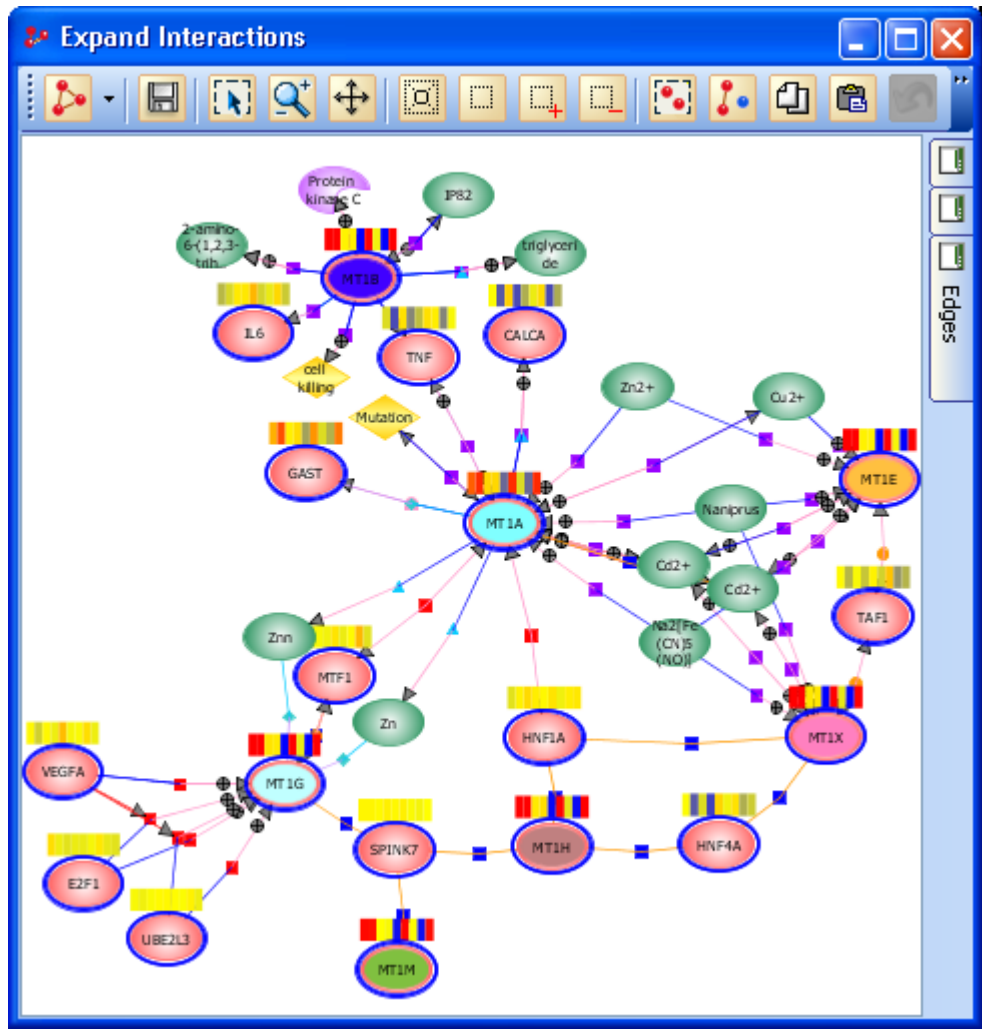


Figure 25.17: Data Overlay

Search can also be performed on the pathway view using the right click dropdown options: ***Search by Properties***. The search wizard described above will be seen.

Explorer:

The saved significant pathways get added to the *Project Navigator* in an hierarchical manner under the experiment. The pathway can be visualized by right clicking on the pathway from the *Project Navigator* section. Right clicking on pathway opens the dialogue box with options like

- Inspect Pathway: It shows the details of the pathway being inspected.
- Open Pathway
- Copy

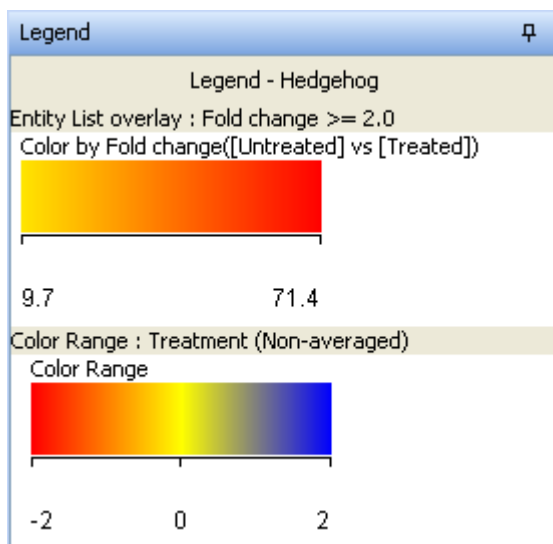


Figure 25.18: Legend for Data Overlay

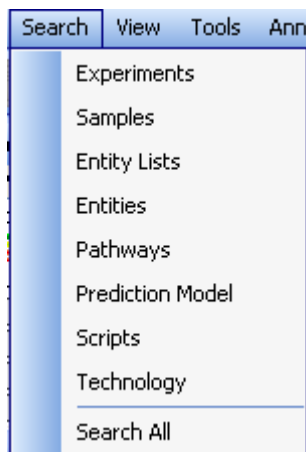


Figure 25.19: Main menu-Search

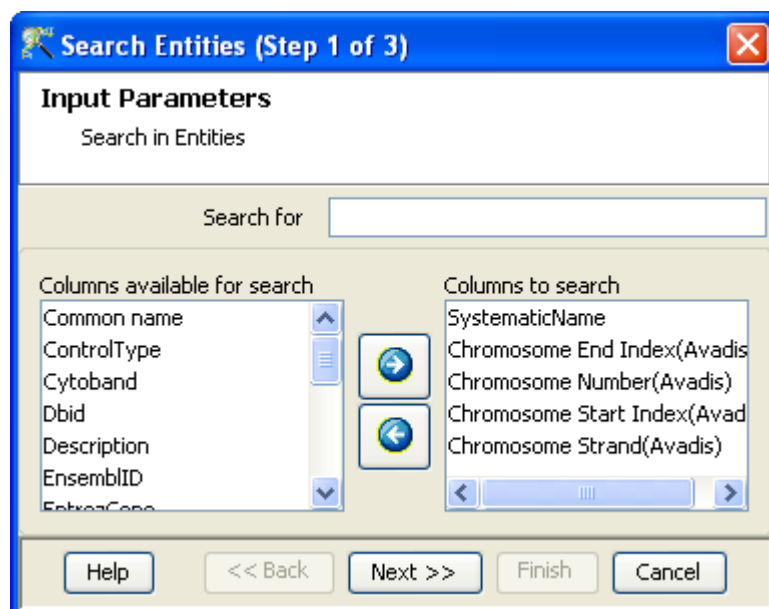


Figure 25.20: Input Parameters

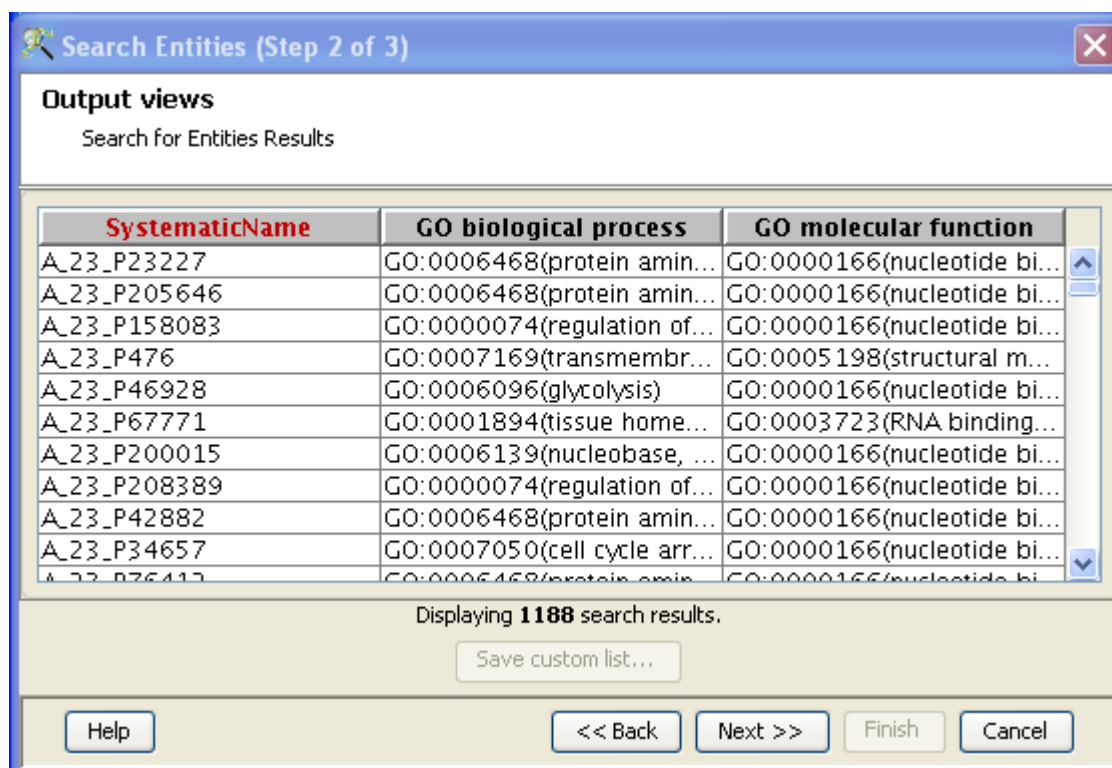


Figure 25.21: Output Views

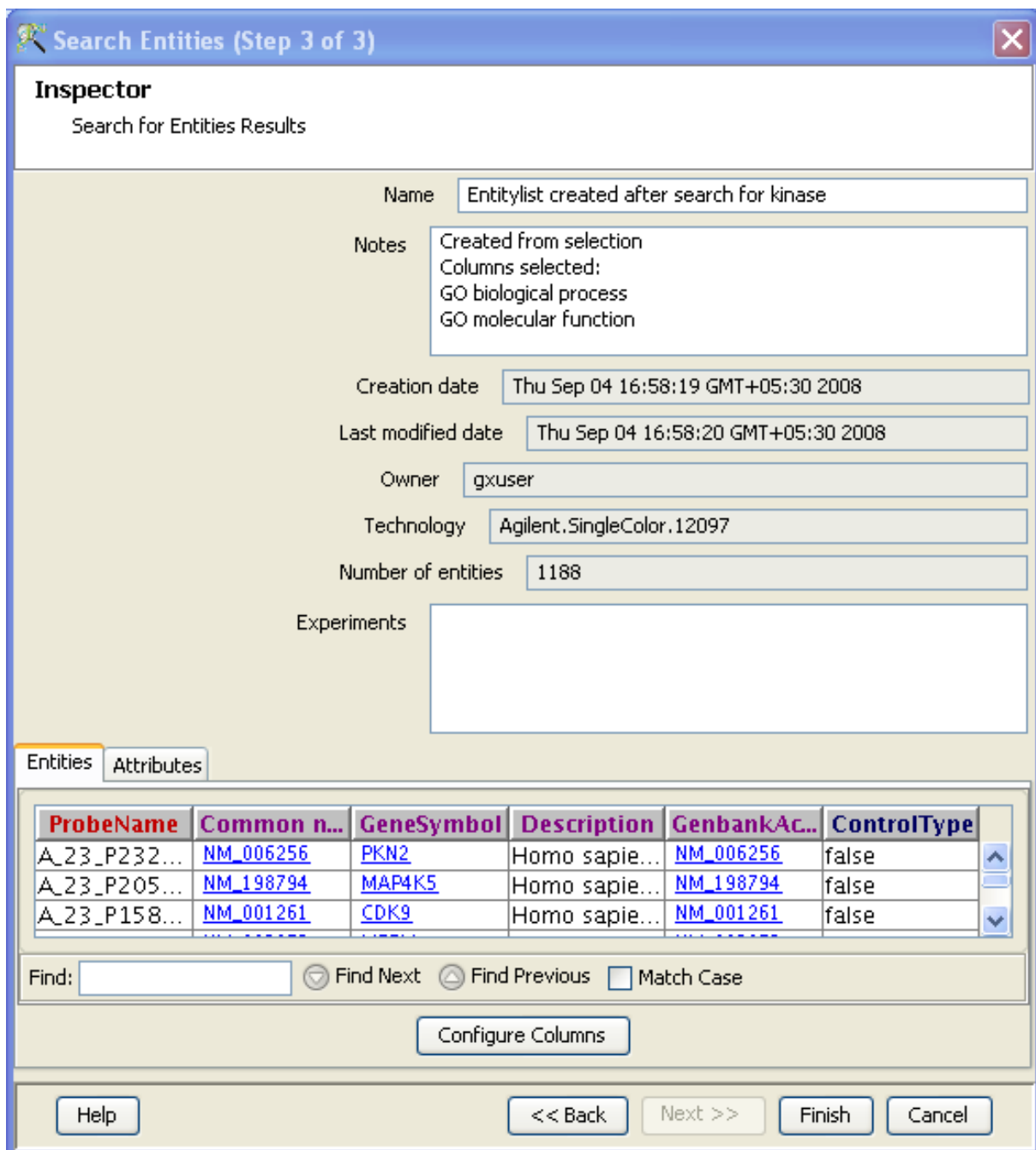


Figure 25.22: Entity Inspector

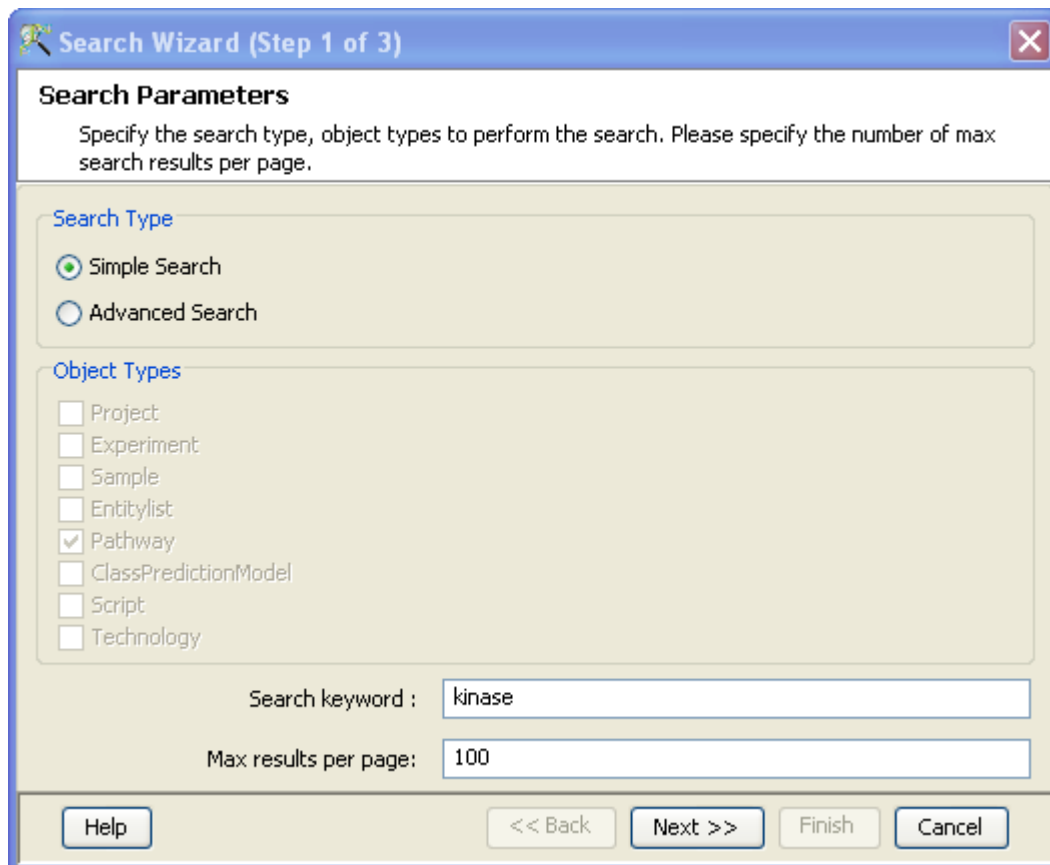


Figure 25.23: Search Parameters



Figure 25.24: Advanced Search Parameters

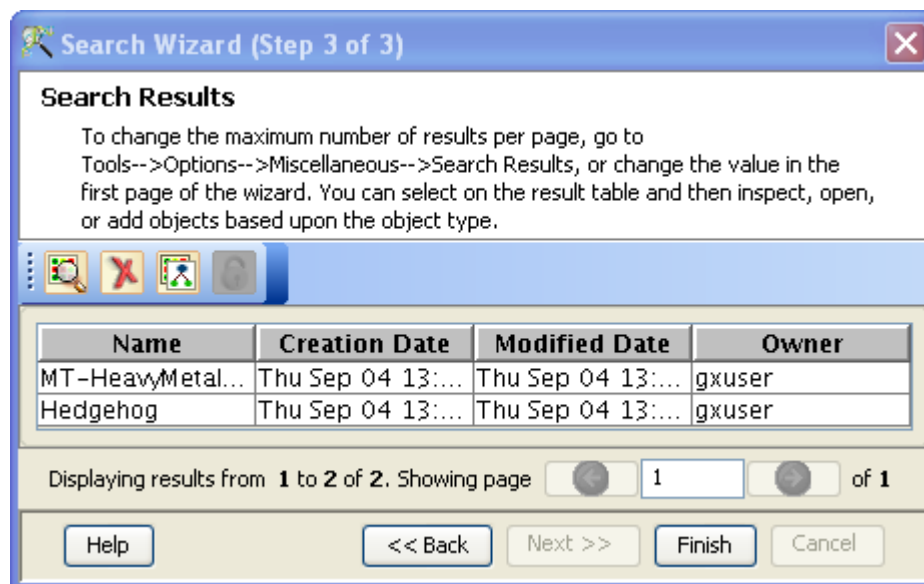



Figure 25.25: Search Results

- Delete Pathway: It permanently removes the pathway from the experiment list.
- Remove Pathway: It removes the pathway from the current experiment list, but the pathway is available for future searches.

25.5.4 Layouts

The view pathway image can be customized according to the users requirements. Any entity or relation can be dragged and moved around the view as per the need. To move multiple entities and relations, use the +Ctrl key or draw a rectangle to select them and drag to move them around.

In addition, **GeneSpring GX** offers a choice of six different pathway layouts. The Layout graph  icon button in the toolbar of the pathway view has several drop-down options for layouts. These are:

- QuadTree FDP
- Neato
- FDP
- Dot
- Twopi
- Cellular

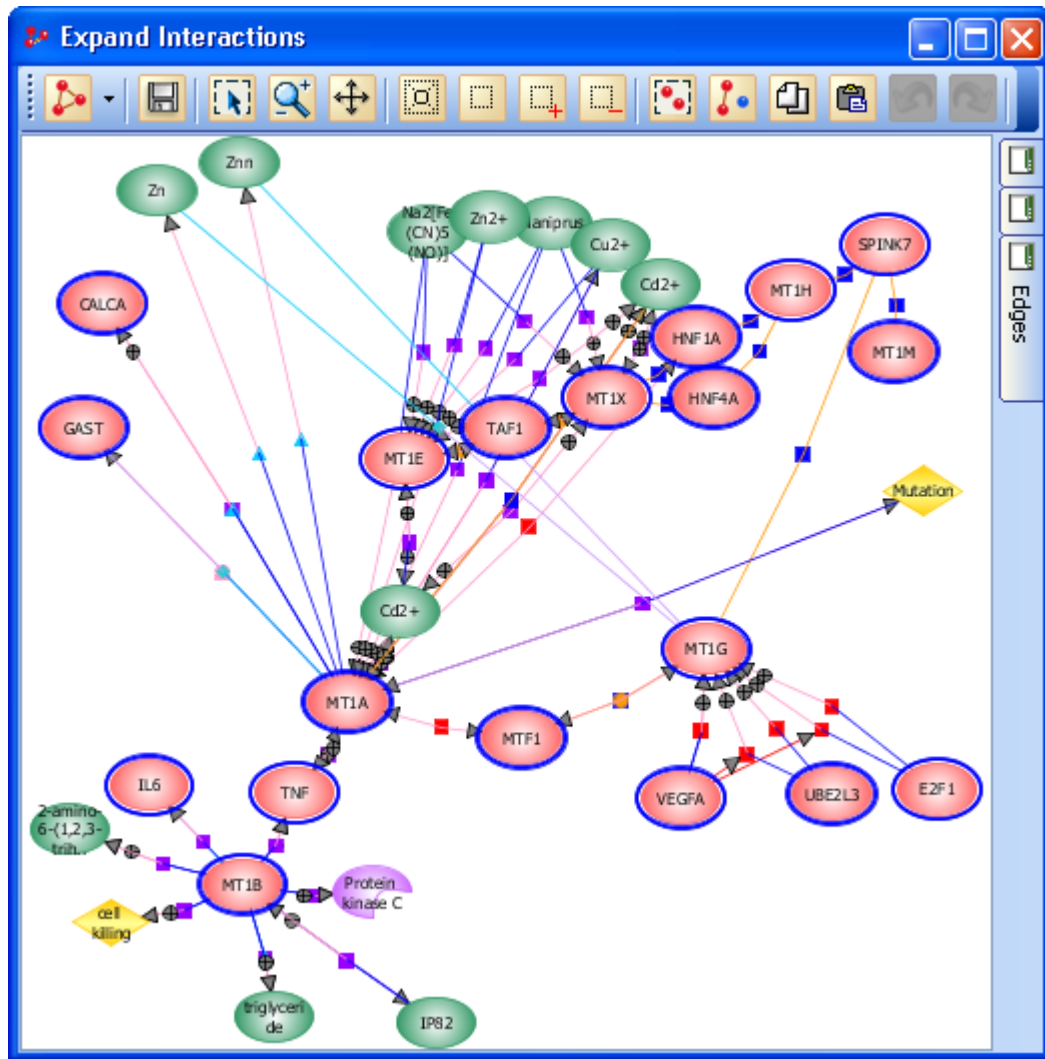


Figure 25.26: Twopi layout

Cellular layout includes the information of the proteins sub-cellular localization within a cell. The information for the proteins cellular localization is derived from the Cellular component of Gene Ontology. Details of Cellular localization is described in section [Database Entities](#). See figure 25.26.

Another useful functionality is creating pathway views is **Merge** selection. A selection of nodes, related by biological function or experimental data can be collapsed into a single node and the user can specify any name for this merged node. All three functions are available as dropdown in the right click menu.

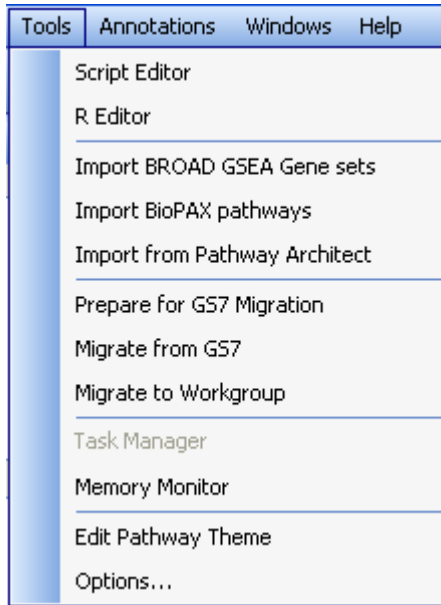


Figure 25.27: Tools→Edit Pathway Theme

25.5.5 Themes

The user can customize the Pathway image. The shapes, colors, fonts and formats of all nodes, relations and edges can be altered. The *Edit Pathway Theme* function can be found in the dropdown options in *Tools*. See figures 25.27, 25.28.

25.6 Extract Relations via NLP

Running Natural Language Processing is another mechanism to create new pathways. NLP can be run directly on PubMed abstracts or on local pdf/doc/html files. NLP will first recognize entities in sentences and then perform information extraction to identify relationships between these entities. The entities NLP can recognize are restricted to those packaged in the pathway relation databases. Additional entities imported into these databases via BioPAX import etc will not participate in NLP. Note again that **GeneSpring GX** determines which organism database to use based on the technology of the currently active experiment.

The natural language processing algorithm that was used to generate the relations database in **GeneSpring GX** can be launched from the UI and made to run on any document of interest. The NLP functionality can be launched in the analysis from *Results Interpretations*→*Extract Relations via NLP*. See figure 25.29

- Step 1 of 4:

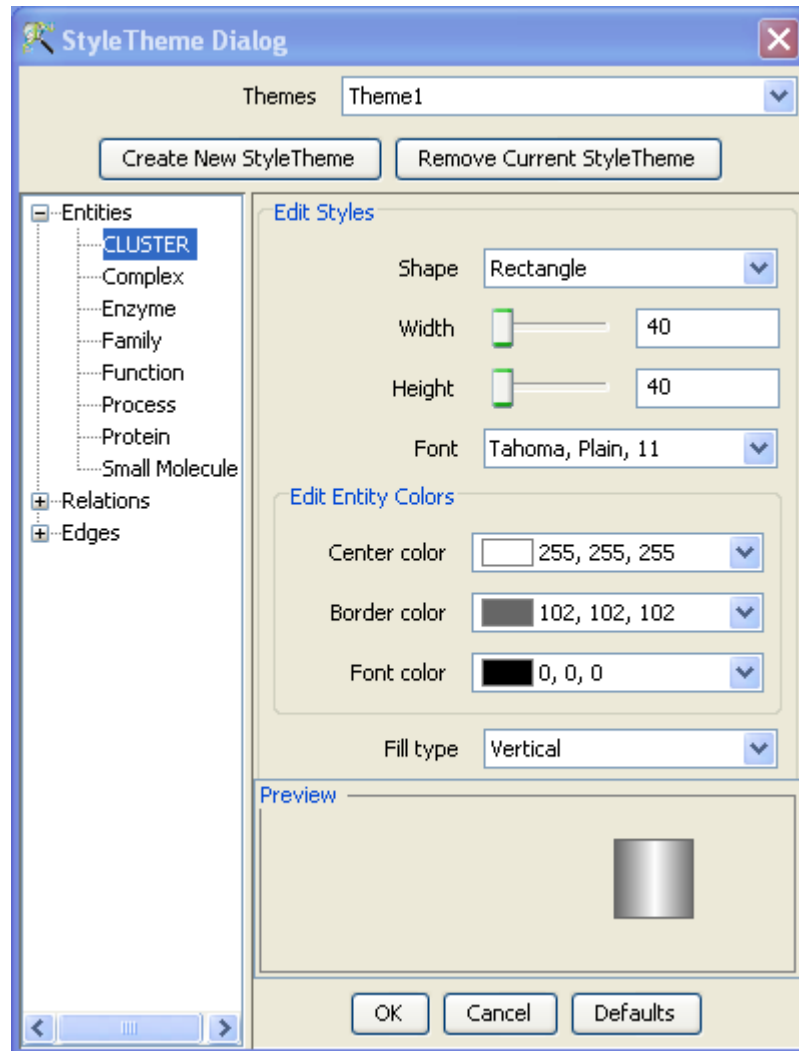


Figure 25.28: Style Theme Dialog

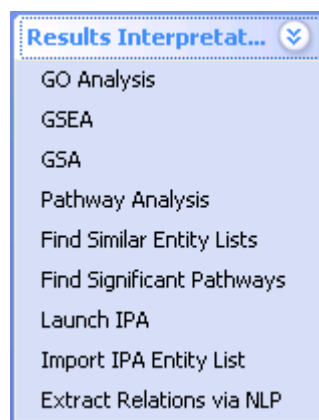


Figure 25.29: Extract Interactions via NLP

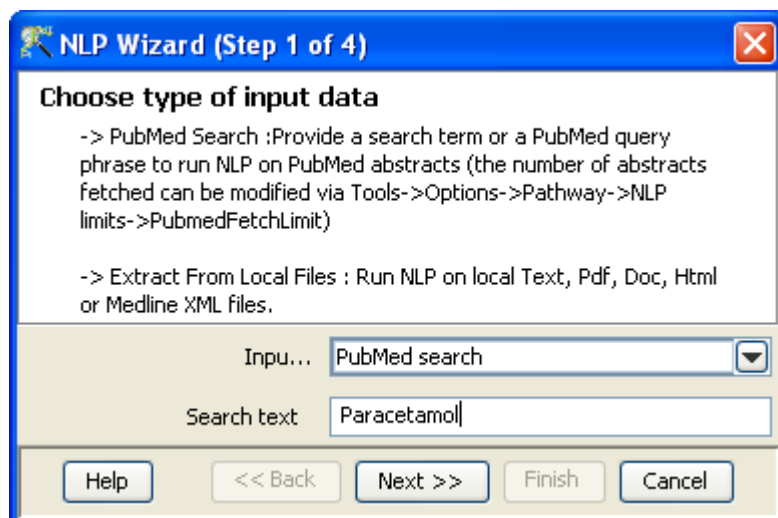


Figure 25.30: Input Data

A wizard launches asking the user to *Choose type of input data*. See Figure 25.30

If the source chosen is “PubMed search”, the user can specify a search query to be submitted to PubMed. The documents returned by PubMed form the basis for the next step. Users also have the choice of running NLP on local files. Doc and pdf files are first converted into text using publicly available converters (antiword and pdftotext respectively), and the text content is then passed on to the next stage.

- **Step 2 of 4:**

The *View tagged content* wizard is launched. The target documents containing the search terms are identified and tagging is performed using the entity dictionary. All molecular and biological process/ function entities present in the **GeneSpring GX** database will be tagged in these searched documents. The color of the tagged entities match the default settings. In the case of PubMed articles (or Medline XML files), the PMID is shown in the left hand column. In all other cases, the name of the file is displayed. Each tagged document can be separately viewed in the right hand column by selecting the corresponding document name or PubMed ID. See Figure 25.31

- **Step 3 of 4:**

All sentences which have at least 2 tagged entities are analyzed by the NLP algorithm. A *Pathway View* is launched with the relations extracted from these sentences. The total number of extracted relations also shown. The sentences from which the relations were derived can be examined (by double-clicking) on the relations. See Figure 25.32

- **Step 4 of 4:**

The *Object Details* wizard is launched. Statistics about the created pathway and its participating entities are provided in a table. The names of the participating entities, the entity type and the Entrez ID of the proteins, wherever applicable are displayed. See Figure 25.33

The user has to click *Finish* to create the pathway. Using the Search pathways option, the pathway can be imported into the experiment and used for *Find Significant Pathways* analysis.

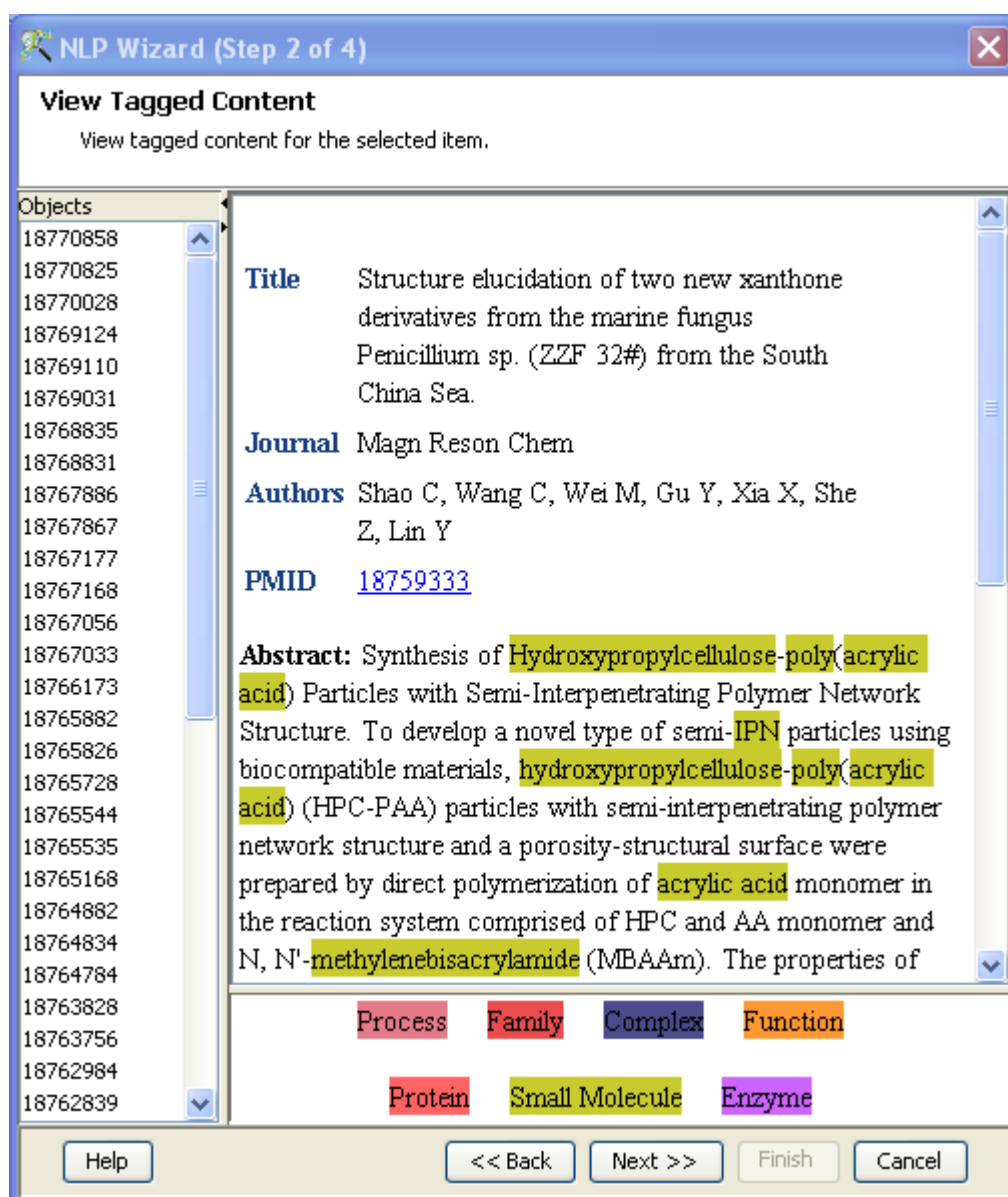


Figure 25.31: View Tagged Content

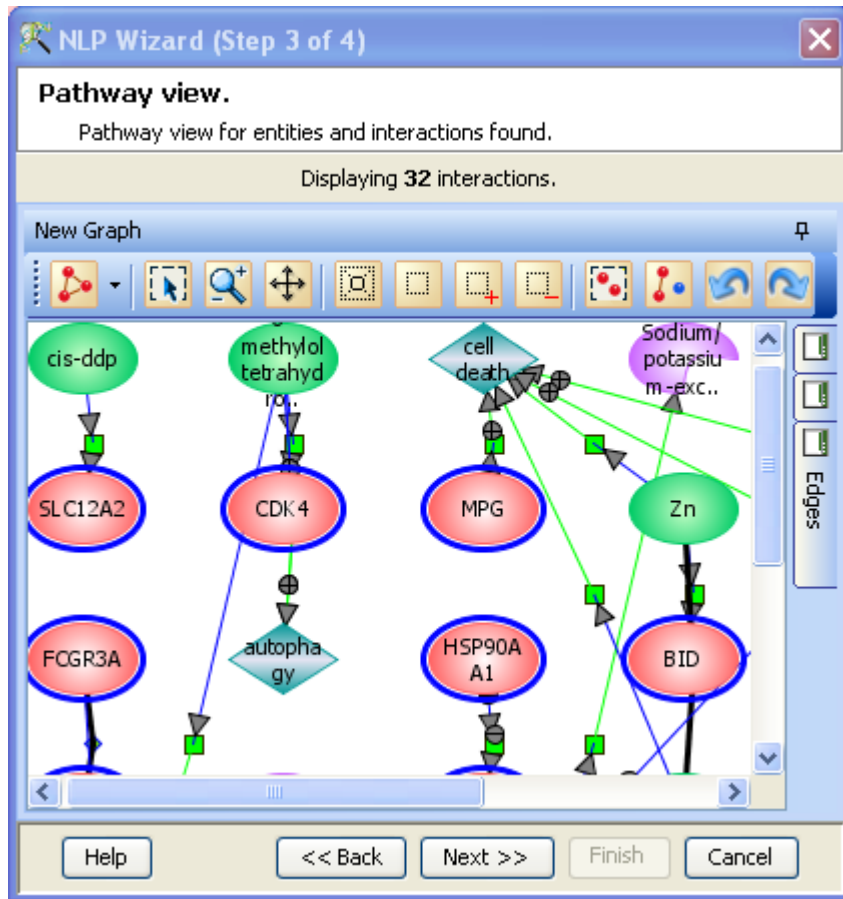


Figure 25.32: Pathway View

25.6.1 NLP Settings

Some of the phases in the NLP process are time and memory intensive. In particular, due to the inherent ambiguities in the English language, and because of looseness in our grammar, it is possible for a sentence to have multiple syntax and semantic parse trees. To prevent the tool from crashing, we have chosen a set of defaults that limit the power of the NLP engine when run through the tool. Some of the differences are:

- smaller lexicon
- limit on the sentence length
- limit on the number of alternate parses that will be processed
- limit on inferencing

At present these settings are not accessible to the user.

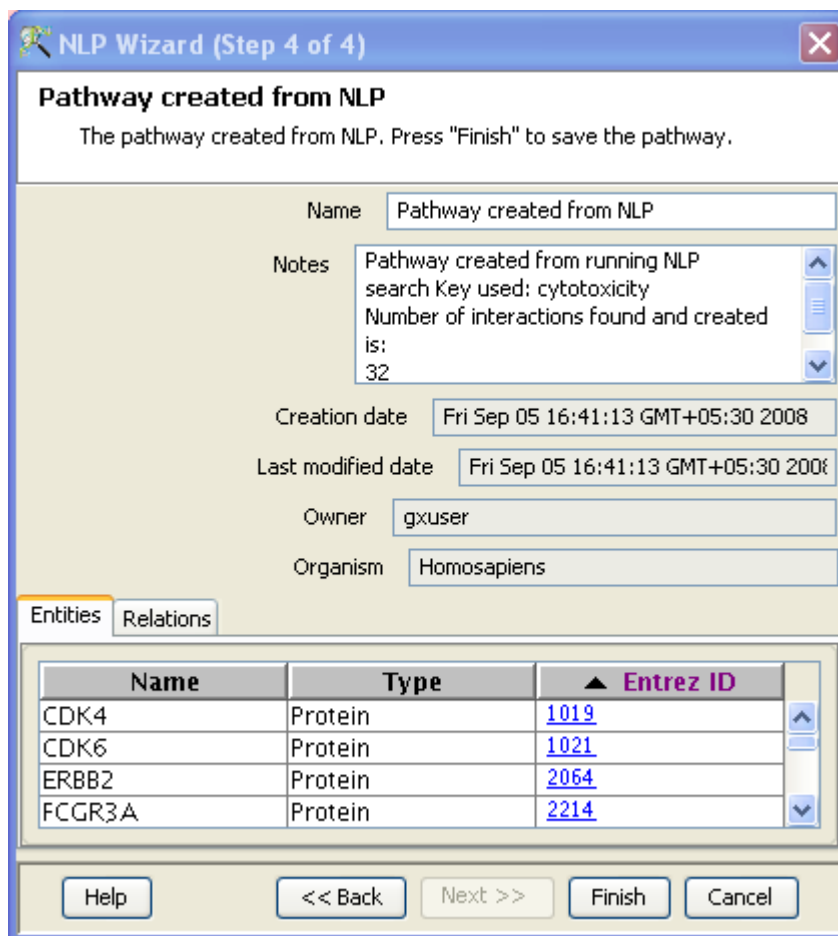


Figure 25.33: Object Details

Another difference with the NLP used to create the databases is that MeSH terms are not used to filter out irrelevant abstracts. Users who wish to do so must construct the PubMed queries appropriately. See the PubMed help for details (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Search_Field_Descr).

25.7 MeSH Pathway Builder

The **MeSH Pathway Builder** in **GeneSpring GX** is a novel way of creating pathways around a *Term* or a *Concept*.

- This functionality works independent of experiments and entities.
- It needs only the 'term/concept' to be input by the user.

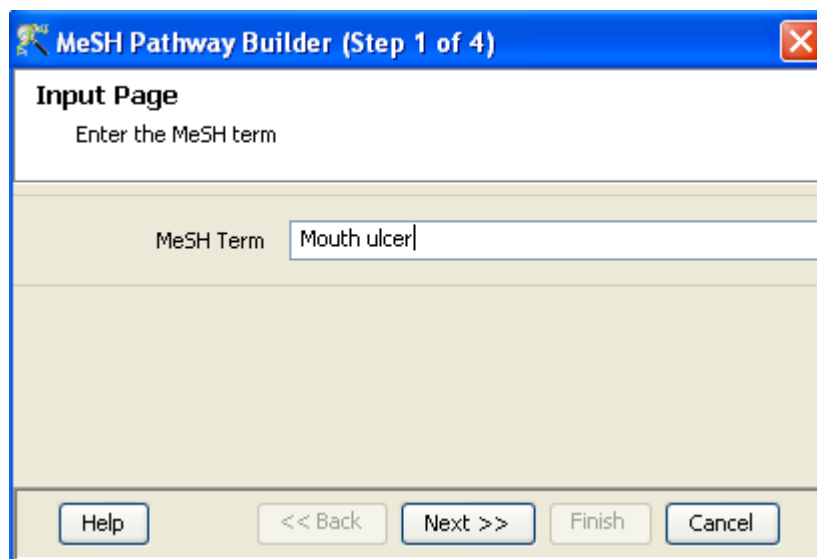


Figure 25.34: Step 1: Input Page

- **GeneSpring GX** pulls out the MeSH terms relevant to this input term/concept from MeSH database (see <http://www.nlm.nih.gov/mesh/meshhome.html>), by exact/substring/synonym matching.
- It obtains all the interactions containing these MeSH terms from the pathway database packaged with **GeneSpring GX**. A list of the relevant MeSH headings along with the number of relations (exact or all relevant interactions) is shown; User can then choose one or more of these to create a pathway with all those interactions.
- User can also filter the interactions based on the frequency of occurrence and the type of relation (direct or indirect) to the 'term'.
- Once the pathway is shown, all the regular operations can be carried out. See [Pathway Viewer](#) for details on the possible operations in pathway viewer.

25.7.1 Launching MeSH Pathway Builder

MeSH Pathway Builder can be launched from the workflow handle *Advanced Analysis* → *MeSH Pathway Builder*. A four-step wizard opens up.

- **Step 1: Input page:** Give the 'Concept' or the 'MeSH' term here. Note that the 'term' need not be technical but a simple phrase or phenomenon of interest will do, eg. Ulcer.
- **Step 2: Select relevant MeSH Terms:** A table is shown listing all the MeSH headings containing the input 'term/concept'. Select one or more MeSH headings and define the filtering options explained below.

Exact Relations	Will include only those interactions which contain the exact MeSH heading that is selected.
All Relevant Relations	Will include all interactions which contain either the exact MeSH heading or the child MeSH terms.
Frequency	Defines the minimum number of Pubmed articles (PMIDs) associated with the MeSH term that an interaction should have. For example, if the <i>Min Frequency</i> is given as 5, the pathway will include only those interactions which have atleast 5 PMIDs that contain the relevant MeSH term. Default is 1.

Table 25.2: Type of relationship and Frequency

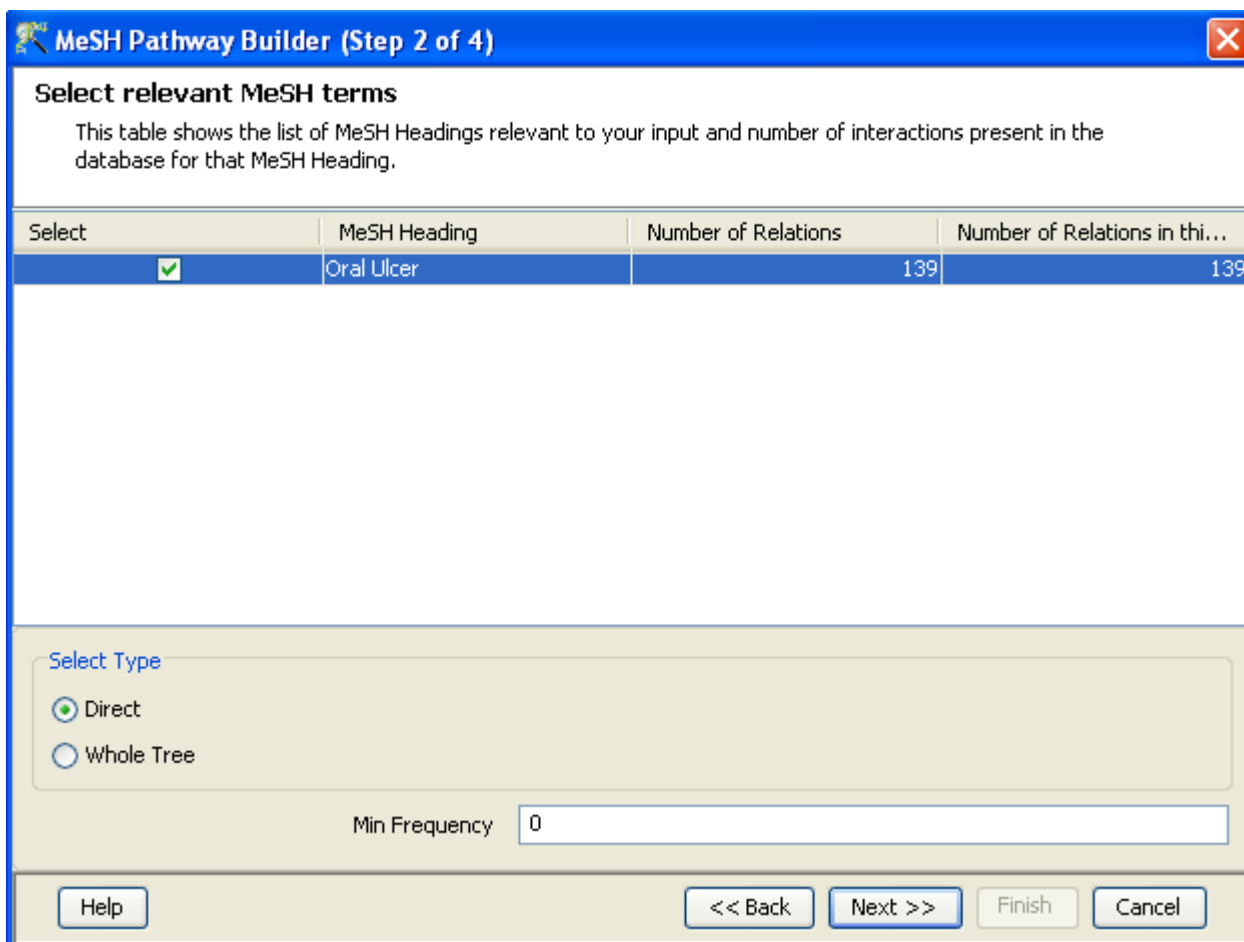


Figure 25.35: Step 2: Select Relevant MeSH Terms

- **Step 3: MeSH Pathway:** A pathway built out of all the interactions associated with the relevant MeSH term is shown here.
- **Step 4: Pathway created from MeSH:** Output is shown here for both entities and relations. Clicking *Finish* will exit the wizard and create a node in the experiment for this pathway.

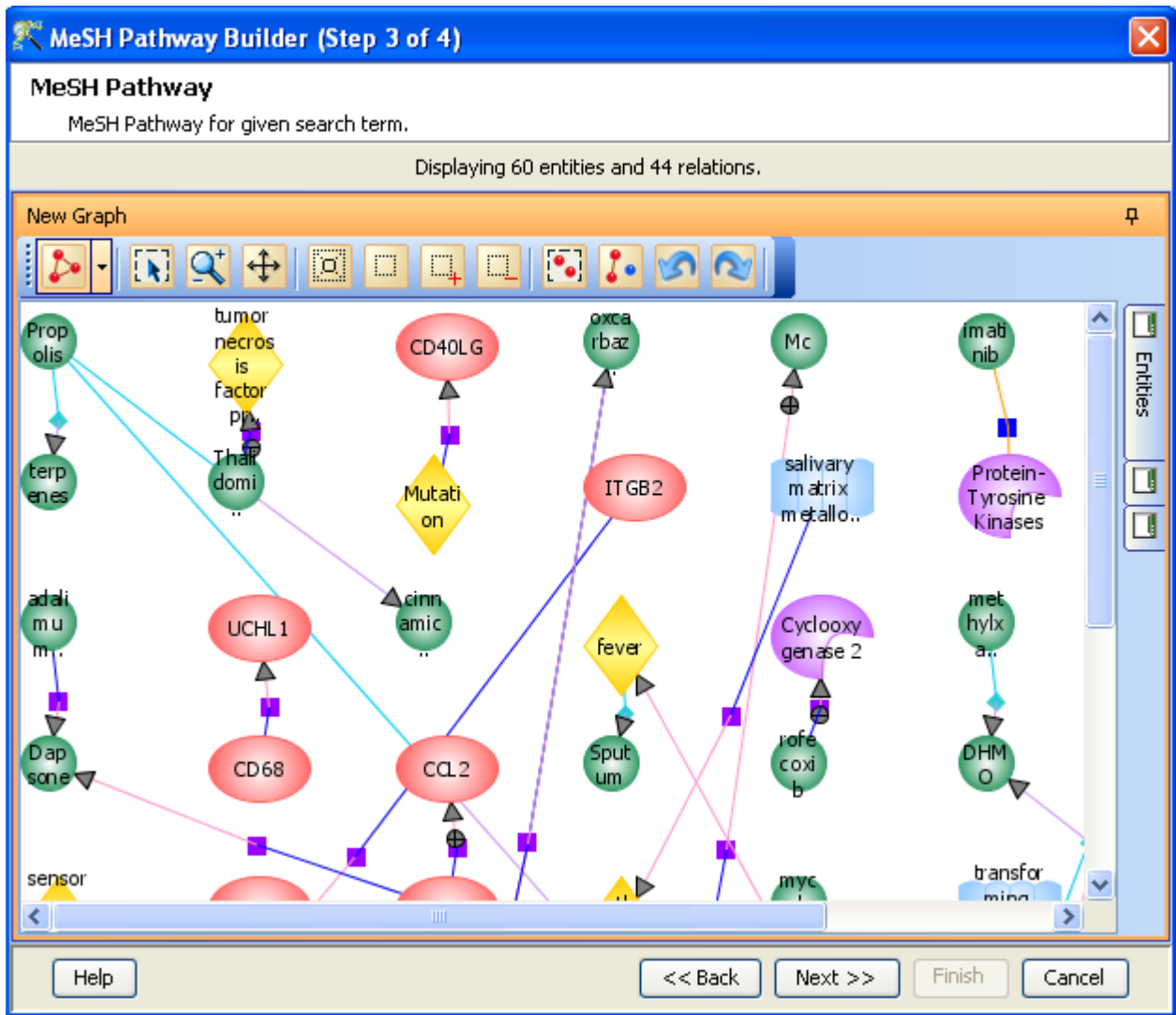


Figure 25.36: Step 3: MeSH Pathway

25.8 Find Significant Pathways

The *Find Significant Pathway* functionality in **GeneSpring GX** helps to identify experimentally proven biological pathways that may be hidden in the user entity list. In other words, this tool allows users to determine in which biological pathways there is a significant enrichment of the genes of interest.

25.8.1 The BioPAX format

GeneSpring GX supports the BioPAX pathways/network exchange format (OWL). Pathways in BioPAX level 2 format are available from a variety of public sources such as KEGG, The Cancer Cell Map, BioCyc

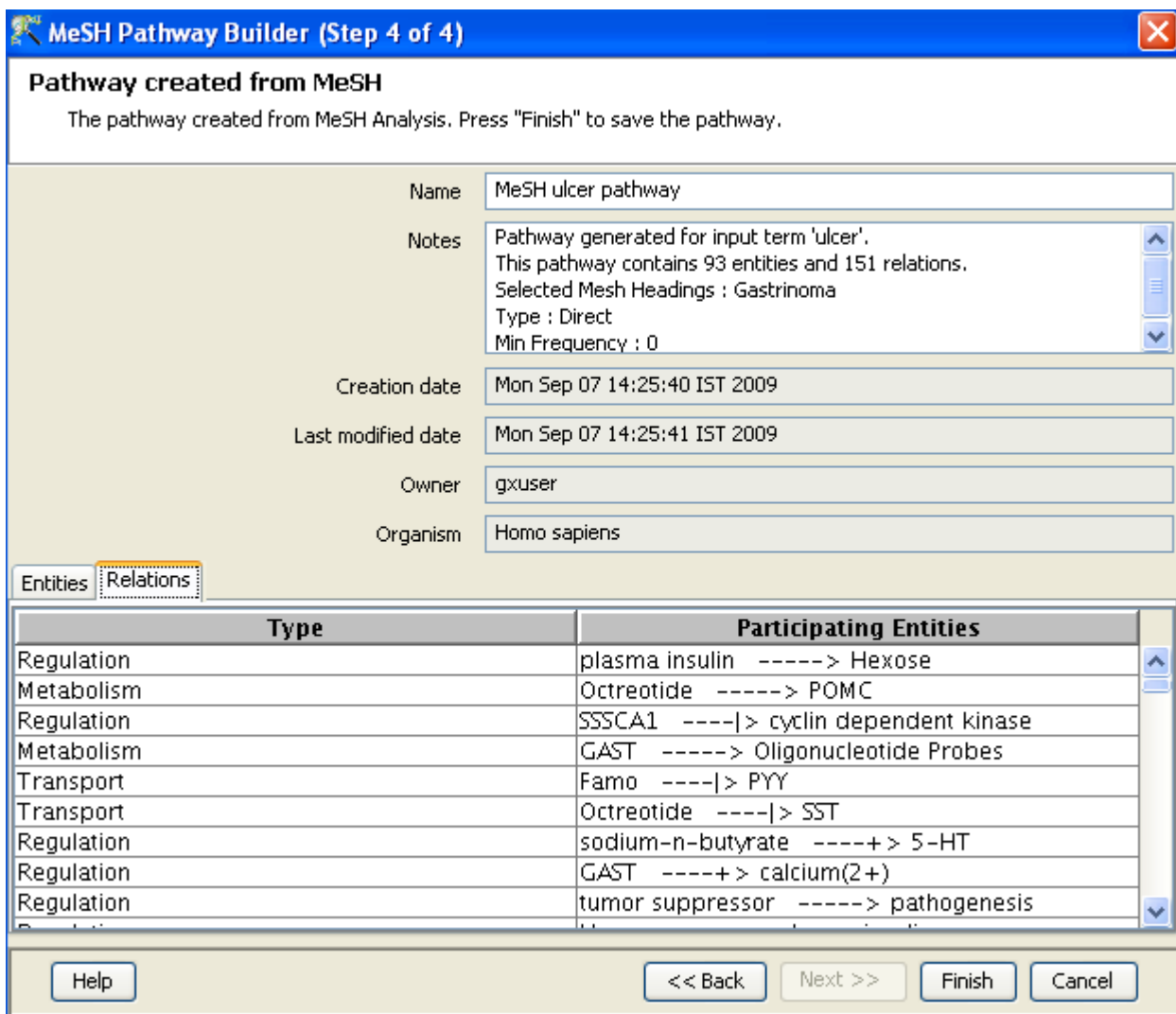


Figure 25.37: Select Pathways to Import

and many others. These can be imported via *Tools* → *Import BioPAX pathways*. The resulting wizard will require that you have downloaded or [created a database for the organism of interest](#) since each pathway needs to be associated with some organism. The list of organisms databases accessible to the tool, appears in the dropdown list. Choose the appropriate organism, select the pathways of interest and press *Ok*; the resulting pathways will be imported and will be accessible via *Search* → *Pathways*. See <http://www.pathguide.org> or <http://biopax.org> for more information on available pathways.

Note: Import of KEGG pathways in the BioPAX format requires non-academic users to obtain a license through the licensor, Pathway Solution, Inc. (pws@kegg.org). Other pathway/networks may require similar license agreements and Agilent Technologies, Inc. cannot be held responsible for unlicensed use of network or pathway data.



Figure 25.38: Choose BioPAX files

The pathways in the BioPAX (OWL) format need to contain the correct annotation information, in order for **GeneSpring GX** to be able to match the proteins in the pathways to the correct entities in the Entity Lists. **GeneSpring GX** uses the Entrez Gene and SwissProt annotation marks to match the genes/proteins to the entities. So it is imperative that both the BioPAX pathways and the technologies for which the pathways are to be used, have the Entrez Gene or SwissProt annotation information. For the Affymetrix, and Illumina technologies, the Entrez Gene is used for matching entity lists with pathways, For Agilent technologies, the SwissProt annotations are used to match entity lists with pathways. For custom technologies, while creating the technology it is necessary to import and mark either Entrez Gene or SwissProt annotations for you to use the pathway functionality. The user needs to download one or more OWL files from these websites to ones local computer.

GeneSpring GX comes pre-loaded with a small set of immune signalling and cancer signalling pathways, courtesy of the Computational Biology Center at Memorial Sloan-Kettering Cancer Center, the Gary Bader's lab at the University of Toronto for the Cancer Cell Map, the Pandey Lab at Johns Hopkins University and the Institute of Bioinformatics (Bangalore, India).

Import BioPAX files

To Import pathway files in the BioPAX format into **GeneSpring GX** the user has to select *Tools*→*Import BioPAX pathways* from the workflow. This launches a wizard **Choose BioPAX files**.

- **Step 1 of 2:**

The user has the option to browse and select the BioPAX files with .owl extensions, to be imported into the experiment. See figure 25.38.

- **Step 2 of 2:**

This is an intermediate step where the user can interactively select the pathways contained within each .owl file. The import wizard automatically parses the BioPAX file and displays the following columns

- *Pathway Name*: Name of the pathway imported
- *File Name*: Name of the .owl file containing the pathway
- *Identified Organism*: organism identified from the pathway file
- *Available Organism*: the selected organism database within the experiment to which the file needs to map.

Pathways where the Identified Organism automatically matches the organism of the experiment (Available Organism) are imported by default. However, if the .owl file does not contain the information for the organisms (Identified Organism column does not contain data), the user can select and import the pathway into any organism database that is available in the experiment. Alternatively, if the user is unsure of which database should hold this pathway, do not select this pathway for import in this session. The user can make a decision later and download the organism database (if present) and then import this pathway file again. If the Identified Organism file has the requisite formation but the database is not currently downloaded into the experiment and the given organism database is available for analysis in **GeneSpring GX**, the user has to go back and download the corresponding database and then import the files again.

A checkbox against each pathway allows the user to make final pathway selection for import. See [Figure 25.39](#)

The imported pathways can be searched using the Pathways menu item in the Search menu. The Find Significant Pathways function will use these pathways to find entities that match between the user defined list and the existing pathways.

25.8.2 Prepackaged Pathways and Migrating Older Pathways

The **GeneSpring GX** samples folder contains a pathways subfolder which in turn contains several subfolders, each corresponding to a different organism. Each such organism folder contains a set of PathwayArchitect .bin files, each containing a hand-curated pathway (either signaling or disease).

These pathways can be imported into **GeneSpring GX** using *Tools* → *Import from PathwayArchitect*. The resulting wizard will require that you have downloaded the database for the organism of interest as described above; this organism will then show in the dropdown carrying organism names in the wizard. Choose this organism, use the select all checkbox to select all pathways and press *Ok*; all the resulting pathways will be imported in and will be accessible via *Search* → *Pathways*.

In addition, note that about 20 human pathways are present in pre-imported form in **GeneSpring GX**. For those users who have an installer version prior to version 10, the pathways present therein are in an

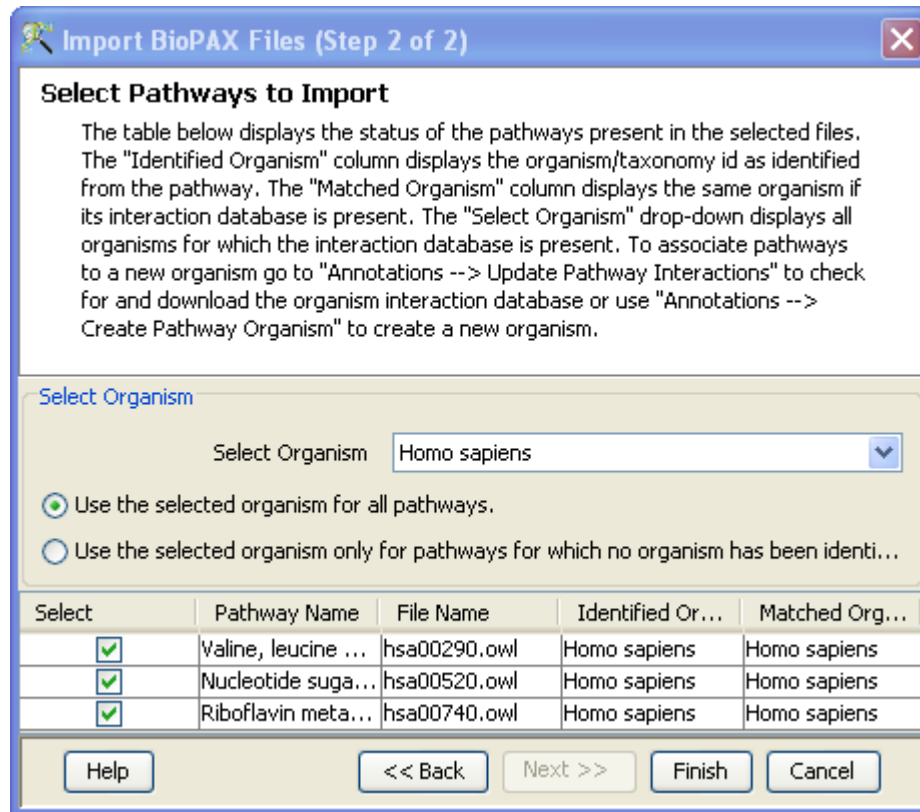


Figure 25.39: Select Pathways to Import

older format that will be obsolete going forward. You can upgrade these pathways to the new format by adding relevant pathways to an experiment, and clicking on these pathways one by one. This will prompt an upgrade of format, which among other things, will register this pathway in the relevant relations database. In the process, the organism of this pathway will be set to that of the technology of this experiment. Note that you should have downloaded the relevant relations databases for these upgrades to happen; in the absence of these databases, old pathways will not open. Old pathways will also not participate in *Find Significant Pathway* queries unless upgraded.

25.8.3 Import from PathwayArchitect

The procedure for importing pathways from PathwayArchitect is akin to that described for pre-packaged PathwayArchitect pathways in Section 25.8.2. This allows the user to import .bin or .xml file into **GeneSpring GX** from the PathwayArchitect tool. The option can be found in the main menu drop downs: **Tools** → **Import from PathwayArchitect**. These pathways can also be added for the *Find Significant Pathways* analysis.

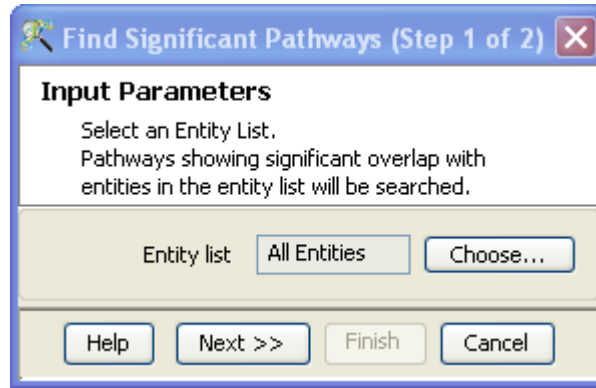


Figure 25.40: Input Parameters

25.8.4 Find Significant Pathways

The collection of pathways saved in **GeneSpring GX** can be searched via *Find Significant Pathway* queries. These queries take an entity list as input and find all pathways which have significant overlap with that entity list. Here, overlap denotes the number of common entities between the list and the pathway. Commonness is determined via the presence of a shared identifier, i.e., Entrez Gene Id, Swissprot Id, or Gene Symbol.

Note that each pathway has an associated organism and a pathway can be searched via **Find Significant Pathways** only if the organism of the active experiment matches the organism of the pathway. The organism of a pathway can be determined by inspecting that pathway and the organism for an experiment can be determined by inspecting its corresponding technology. Note that the former is non-editable while the latter is indeed editable and can be modified in instances where there is a mismatch. The **Find Significant Pathways** function can be launched from *Workflow*→*Results Interpretation*→*Find Significant Pathways*. A wizard appears to guide you through the analysis.

Note: If the starting point is an Entity list, then the results include a p-value column whose computation implicitly makes use of the “All Entities” list. The p-value column is absent in the context of Pathway Experiments, since there is no all-encompassing list of entities.

- **Input Parameters-Step 1 of 2:**

The only input required for the **Find Significant Pathways** Analysis is the entity list of interest. The user would be able to determine whether there is any significant overlap with already known pathways. By default, the active entity list in the experiment is chosen. To change the entity list, click on the *Choose* button and select an entity list from the tree of entity lists shown in the window. See figure 25.40.

- **Viewing and Saving the Results-Step 2 of 2:**

This view of the wizard shows a number of significant pathways satisfying the p-value cut-off (default is $p=0.05$). The resulting pathways are shown in two windows. The window on the left shows

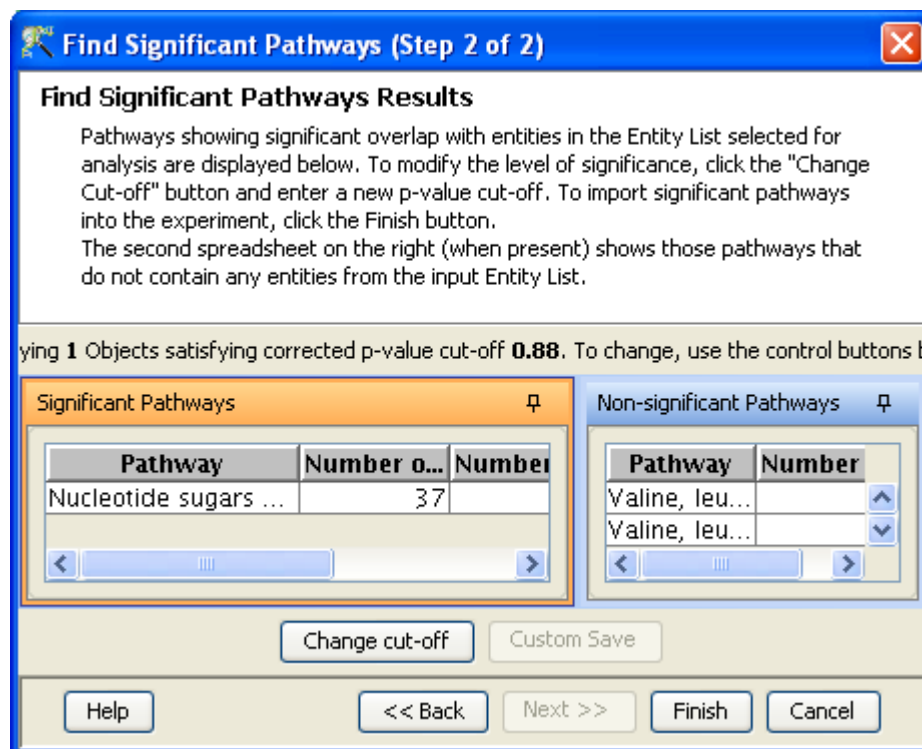


Figure 25.41: Results Window

Significant Pathways (i.e., those which have an entity in common with the input entity list) and the window on the right shows **Non-Significant Pathways** (i.e., those which do not have any entity in common with the input entity list).

For the resulting significant pathway window on left, the *Pathway Name*, *Number of Entities, Matched with technology*, *Matched with Entity List* and *p-values* are displayed.

For the **Non-Significant pathways**, *Pathway Name* and *Number of Entities* are displayed. Overlap of entities is determined based upon the presence of either an Entrez ID or a SwissProt ID in the relevant pathway. Once the number of common entities is determined, the p-value computation works exactly as in the case of **Find Similar Entity Lists** using the equation 18.1 (this is effectively the Hypergeometric method or the Fisher's exact test). To modify the level of significance, click on the *Change cut-off* button and enter a new p-value cut-off. The spreadsheet of results will be automatically updated with respect to the new p-value cut-off. See figure 25.41

To save all the significant pathways in the experiment, click on the *Finish* button. To save a subset of the significant pathways, select the same and click on the *Custom Save* button.

25.9 Pathway Experiment

GeneSpring GX offers separate pathway analysis experiment, independent of the microarray analysis workflow. This feature aims to provide a wider use of pathway functionalities to cover different types of data. The input to a pathway experiment analysis is a tab-delimited text file. The file should contain a column with standard public IDs for the entities (Entrez gene, UniGene or gene symbol for proteins), associated data columns for all samples eg. signal intensity values for every sample and list-associated values such as fold change or p-values from standard statistical tests. Only the ID column is mandatory in the file, the rest of the columns are useful to visualize the pathway in the context of the data using data overlay. Different types of data can be imported into **GeneSpring GX** which may include:

1. Genomic
2. Proteomic
3. Metabolomic
4. Combinations of proteins and small molecules
5. Genes corresponding to interesting biological processes and functions defined by gene ontology.

25.9.1 Launching a Pathway Experiment

When **GeneSpring GX** is launched, the user is provided with the options to create new experiments. Amongst the different types of experiments displayed in the dropdown menu options, the user can choose to create a new **Pathway Experiment**. User can also select “Pathway Experiment” from the drop down options from **Projects** in the main menu. See figure [25.42](#)

Data Import

- **Step 1 of 4:** As in all other **GeneSpring GX** experiments, the user is prompted to provide the Input parameters which include a name to the technology associated with the experiment and specify an organism from the drop down menu. In this first step, the user is prompted to choose a data file for upload which contains a list of the entities for pathway analysis. See figure [25.43](#)
- **Step 2 of 4:** This step prompts the user to choose an Identifier column. By default the first column is selected. **GeneSpring GX** performs automated ID match. The available IDs are present in the “Available Items” box and the matched ID is displayed in the “Selected items” box. A mix of IDs for different entity types is permitted. **GeneSpring GX** recognizes all standard public IDs for protein entities (Entrez, Nucleotide, UniGene and gene Symbol). All associated numerical data columns are imported with the entity list.

See figure [25.44](#)

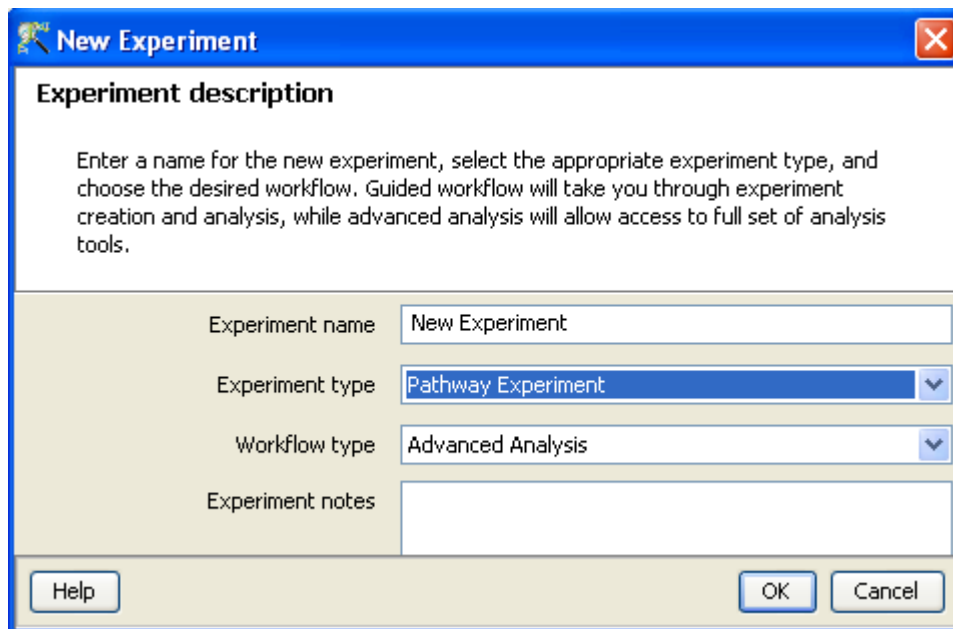


Figure 25.42: Pathway Experiment Creation

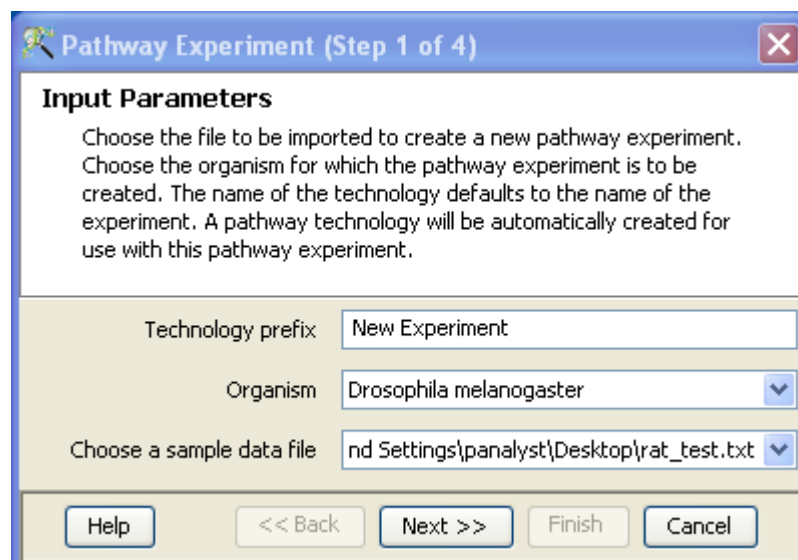


Figure 25.43: Input Parameters

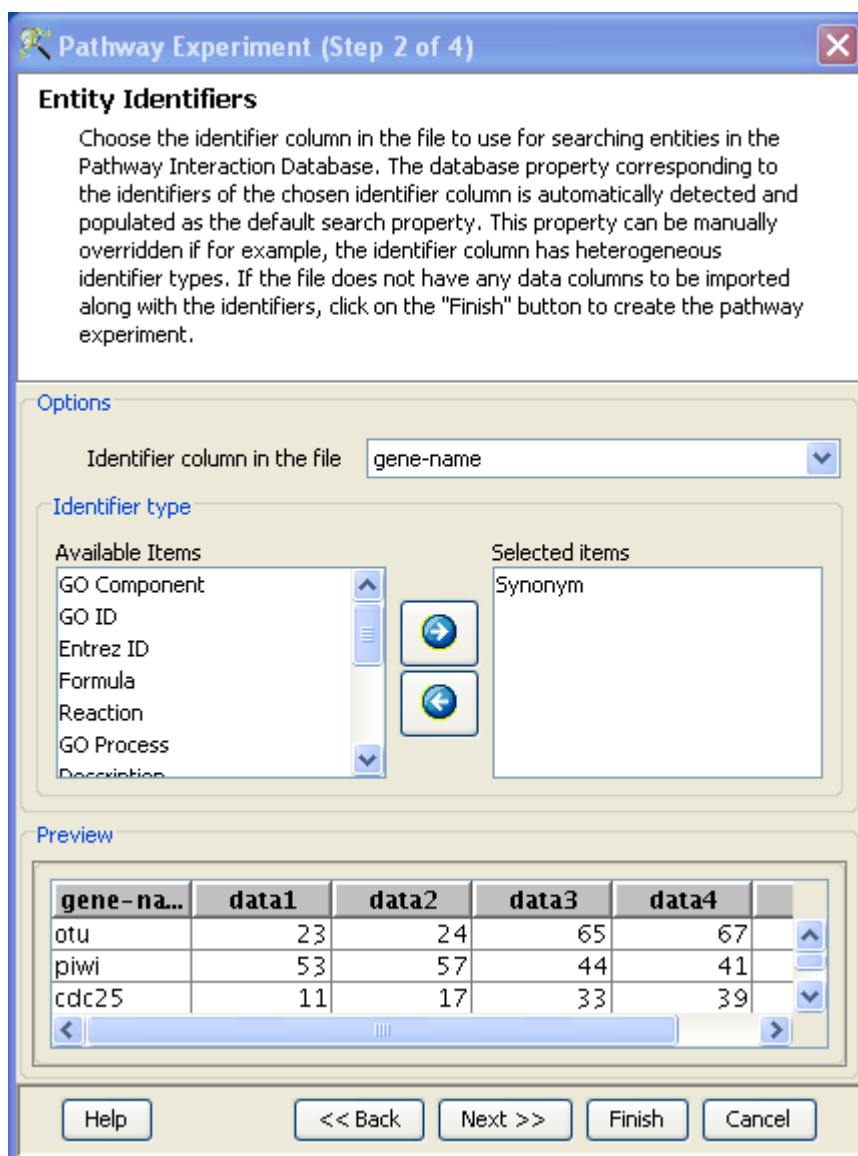


Figure 25.44: Import List from File

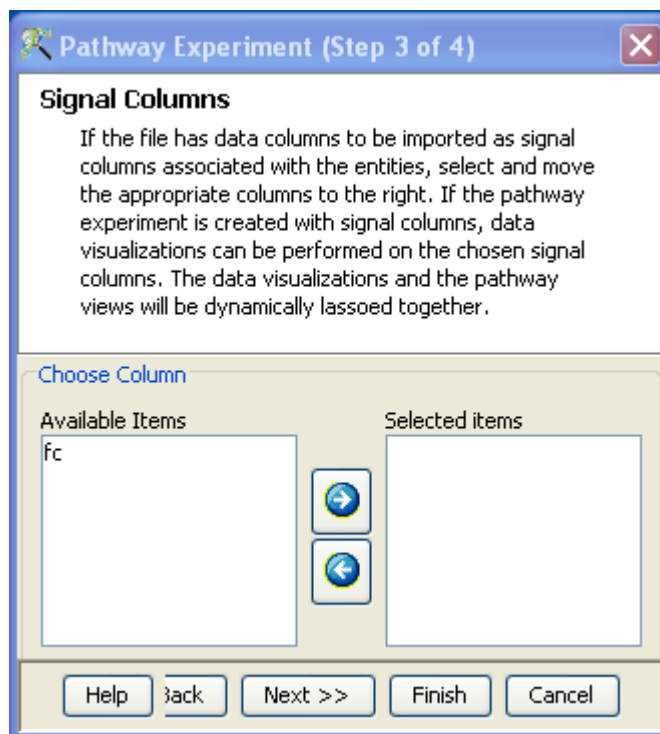


Figure 25.45: Choose signal columns

- **Step 3 of 4:** The wizard **Input parameters - Choose signal column** allows the user to select any columns containing real numbers as sample data columns. This step is optional. If the user does not have any data column, then one can click finish and go directly to the entity inspector. If multiple rows in the input file map to the same entity in the database, the average of the data rows is taken.

See figure [25.45](#)

- **Step 4 of 4:** This wizard **Choose extra column** allows the user to select list associated data columns etc. If multiple rows contain the same entity, data in the first row is chosen. This step is also optional.

See figure [25.46](#)

Spreadsheet: The wizard displays a spreadsheet matching all entities from the database search. A list for **All Entities** appears in the Explorer. Subsequent steps of Pathway Analysis can be run on this list.

Workflow - Pathway Experiment

A typical **GeneSpring GX** experiment is provided with a recommended workflow to guide users through the analysis. A pathway experiment workflow consists of the following steps:

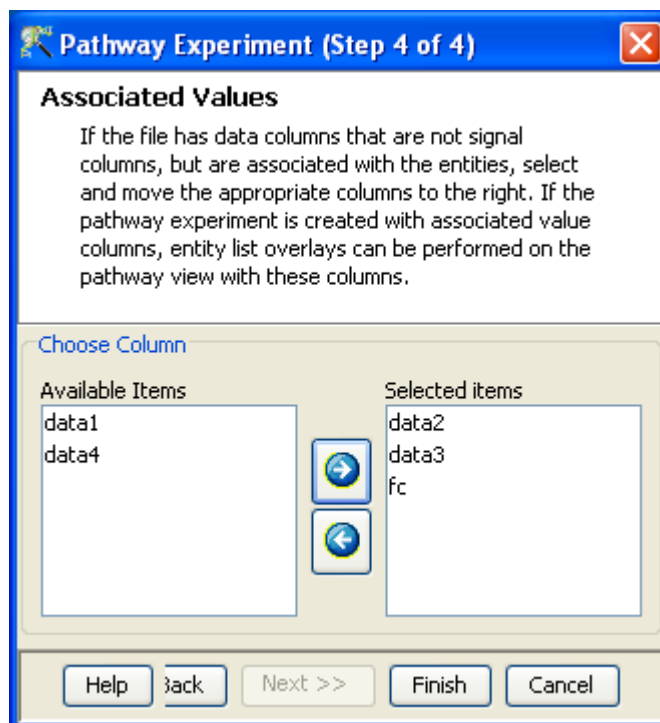


Figure 25.46: Choose extra column

1. **Experiment Setup:** This section prompts the user to group the samples using "Experiment Grouping" (sample 1,2,3 belong to groupA and sample 4,5,6 belong to group B) and also specify the "Interpretation" (compare group A to group B). This grouping and interpretation information will be subsequently used interpreting the overlaid data.
2. **Simple (guided) Analysis:** A two step guided analysis to create biological pathways. Default network creation settings are used.
3. **Advanced Analysis:** In this analysis, the user can specify filters on entities, relations, and their associated properties, to build the pathway networks. There is also an option to give a term and create a pathway with interactions associated with this term. See [MeSH Pathway Builder](#) for details.

See figure 25.47.

25.9.2 Lassoing

Pathway operations can be run from two places - the workflow browser and the pathway view. A new pathway view can be launched from the toolbar. Workflow operations always run off the selected entity list in the Navigator. Operations launched via the views's right click menu will run off the selection in the view.



Figure 25.47: Pathway Experiment Workflow

Selected entities in one pathway view will be selected in all other open pathway views within the same experiment. However, not all selected entities in the pathway views will be selected in spreadsheets and other data views (like scatter-plot). In particular, this lassoing is based on Entrez IDs, and so this selection is limited to entities which have associated Entrez IDs. Clicking on an entity list will highlight corresponding objects on the pathway view provided they have a matching Entrez ID. Clicking on a list will restrict all other data views to just entities contained in the chosen list.

The highlighting and restriction effects of clicking on an entity list will also carry over to other open experiments in the same projects. Again this is done based on Entrez IDs and behaves as described above.

25.9.3 Simple Analysis

The Simple Analysis guides the user to perform a cursory pathway analysis in two steps: Default parameters are set for the network building algorithms. Common filters that are biologically meaningful to users have been included in this section of the Workflow. These are:

- Direct Interactions Network
- Targets and Regulators
- Network Targets
- Network Regulators
- Network Binders
- Post-Translational Modifications (same as network modifiers)
- Transcriptional Regulation
- Transport regulation
- Metabolism regulation
- Small Molecules
- Biological Processes
- Shortest Connect

Shortest Connect finds the smallest uninterrupted number of relations to connect a set of entities.

For details on the functionalities, refer to the section on [Simple Analysis](#) in the Pathway Analysis within Microarray Experiment section.

25.9.4 Advanced Analysis

The advanced analysis section allows the GS user to customize all the algorithms and pathway functions. The workflow consists of the following utilities:

- Pathway Analysis described above in [Pathway Analysis within Microarray Experiment](#)
- Extract Pathways via NLP described above in [Extract Relations via NLP](#)
- MeSH Pathway Builder described above in [MeSH Pathway Builder](#)
- Find Significant Pathways described above in [Find significant Pathways](#)

25.9.5 Exporting Pathways

GeneSpring GX users can export their pathway experiment data in several ways: Exporting Pathway images is a common requirement of the analysis. In **GeneSpring GX**, two options are provided for export of pathway images:

- *Image*: The user can save an image of the current pathway view as a .png or .jpg and custom set the resolutions and image sizes.
- *Navigable HTML*: A navigable HTML image is created. The user can open the image in a browser and click on any node or relation to see its associated properties.

These options can be selected from the right click drop-down menu Export As→Image or Navigable HTML.

See figure [25.48](#)

For a given pathway the list of participating entities and relations can be exported as follows:

- Right click on a given pathway image. Select **Show Report**→**Entities Table** (or **Selected Entities Table**) to generate a list of the entities. Right click on the spreadsheet to export a .txt file.
- Right click on a given pathway image. Select **Show Report**→**Relations Table** (or **Selected Relations Table**) to generate a list of the relations. Right click on the spreadsheet to export a .txt file.

The user can customize the column selections by right clicking the spreadsheet, selecting Properties and configuring the columns.

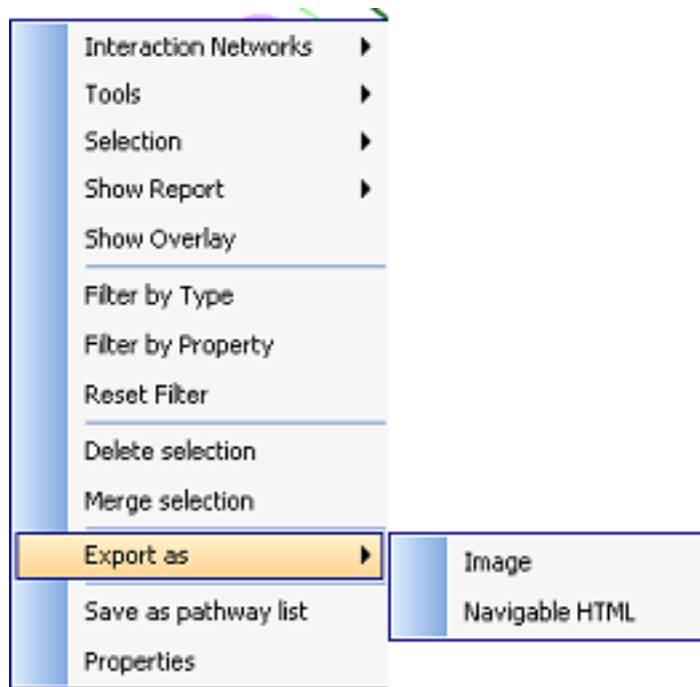


Figure 25.48: Data Export

Transferring Pathways

Pathways can be transferred from one installation of **GeneSpring GX** to another as part of *project-zips*, which carry all data in selected experiments in a project. To create a project-zip of the current project use *File*→*Export Project Zip*. When imported into another instance of **GeneSpring GX** via *File*→*Import Project Zip* all pathways will also be imported. However, when clicked for opening, such pathways will be re-registered with the relations database associated with this instance of **GeneSpring GX**.

25.10 Pathway Database

The Pathway module of **GeneSpring GX** is seamlessly integrated with a database of relations between various biological molecules and processes. The molecules and processes are depicted as entities and their biological interactions as relations. In a pathway view entities form the nodes of the graph and the edges depict the relations.

By default, the following databases can be updated from the menu bar *Annotations*→*Update Interactions*.

- *Homo sapiens*

- *Mus musculus*
- *Rattus norvegicus*
- *Drosophila*
- *C. elegans*
- *Saccharomyces cerevisiae*
- *Escherichia coli*
- *Arabidopsis thaliana*

For other organisms databases, the user can do the following:

- Create a database from **Annotations**→**Create Pathway Organism**. This database can then be populated with appropriate BioPAX pathways (if available) of the organism and **Find Significant Pathways** can be performed on this database.
- For availability of a custom NLP database for a particular organism, contact `informatics_support@agilent.com`.

For additional details, refer to section on [Working with Other Organisms](#).

25.10.1 Pathway Database Organization Overview

Each organism database comprises the following entities: proteins, small molecules, processes, functions, enzymes, complexes and families. Of these only proteins are organism specific while the others are largely organism independent. With this in mind and for reasons of efficiency, the interactions database is organized in a hierarchical fashion. This could be an arbitrary hierarchy in general, but in practice it has 2 levels. The top level is termed *Generic* and contains all non-protein information that is common across organisms. The second level comprises the various organism specific entities (largely proteins) and relations specific to the organism.

This organization is also reflected in the manner in which NLP was run on Medline abstracts. From the overall MeSH hierarchy, about 3000 species terms were identified. These terms were used to divide the approximately 16 million Medline abstracts available at the end of 2007 into two groups: those which had a species MeSH term and those which did not. The latter set of Medline abstracts were marked as “Generic” and the non-protein relations arising from this set of abstracts is shared by all the organisms.

Organism specific relations arise from:

- abstracts that have MeSH terms relevant to the organism

- relations from “Generic” abstracts that have protein participants which can be mapped to protein entities of the organism.

Each organism specific database itself has two sections. A systems section and a user section. All prepackaged entities and relationships reside in the systems section while any new entities and relationships added by the user (via BioPAX import, say) are added to the user section. Any pathway construction or entity matching query will query both systems and user sections as well as the Generic database to provide a complete answer to the query.

25.10.2 Database Entities

The Pathway database contains the following basic types of entities:

- **Complexes:** This class of molecules consists of more than one proteins that physically bind each other and are biologically active and stable, in their combined form.
- **Enzymes:** Proteins acting as biocatalysts in a metabolic reaction. These entities are particularly important in depicting a biochemical network.
- **Family:** A group of proteins related by structure, function or some biological parameter.
- **Function:** A class of molecular functions. Eg. Protein kinase activity.
- **Process:** A class of known biological processes. Eg. Apoptosis
- **Protein:** Represents genes, their transcripts and protein products for a given organism.
- **Small Molecules:** includes classes of molecules like drugs, metabolites, metals etc.

The visual representation of the entities and relations in a pathway view is user configurable from *Tools*→*Edit Pathway Theme*. The default representation is depicted in Figure 25.10. Details on how each entity list was created are listed in the section below titled

This section describes how the Pathway entities and relations were created and their properties were obtained. **GeneSpring GX** supports pathway analysis for the following organisms:

- **Proteins:** Protein entities were obtained by parsing Entrez Gene data available from NCBI for each organism. Protein entities will differ amongst the organisms. All the entities from different species of *Drosophila*, *E.coli* etc. are present in a single database. The available properties for proteins are - Synonym, Description, Cellular Localization, Entrez ID, GO ID, GO Process/Function/Component, Note, Nucleotide, Protein (accessions), and UniGene.

Cellular Localization The following 8 GO Cellular Component terms form the “Cellular Localization” lexicon.

- Plasma membrane
- Cytoplasm
- Nucleus
- Extracellular matrix
- Mitochondrion
- Chloroplast (plants only)
- Golgi apparatus
- Endoplasmic reticulum

Distance of these terms to each of their descendant terms in the GO hierarchy is computed. Proteins which have GO Cellular Component annotations are assigned a single cellular localization term based on this distance computation.

- **Small Molecules**

The list of small molecules was created using four primary sources:

- MeSH 2008 descriptor records
- MeSH 2008 supplementary concept records
- ChEBI
- PubChem

Details on how the sources were used are given below.

MeSH substance descriptor records: Each MeSH record is associated with a set of concepts and semantic types. The records were limited to the following semantic types:

- Antibiotic
- Biologically Active Substance
- Body Substance
- Carbohydrate
- Chemical Viewed Functionally
- Chemical Viewed Structurally
- Chemical
- Eicosanoid
- Element, Ion, or Isotope
- Hazardous or Poisonous Substance
- Hormone
- Immunologic Factor
- Indicator, Reagent, or Diagnostic Aid
- Inorganic Chemical
- Lipid
- Neuroreactive Substance or Biogenic Amine
- Organic Chemical
- Organophosphorus Compound

- Pharmacologic Substance
- Steroid
- Substance
- Vitamin

Some filtering was carried out to remove records likely to be only proteins. For each record the name was set as the MeSH descriptor name, while all the terms listed in the concepts section were stored as aliases. In addition, Registry numbers (if any) were stored as CAS property values. MeSH supplementary concept records were parsed in a similar manner.

PubChem: PubChem records having CAS registry numbers that were mentioned in Medline annotations were retrieved.

ChEBI: Data from ChEBI (Chemical Entities of Biological Interest) was incorporated.

To limit the size of the final database, we consider only those records that contain terms participating in NLP derived relations. Because of the overlap between the data sources, it is possible that the same term is present in multiple records. Filtering of these records providing “useful” terms was carried out as follows:

- “Generic” records which provide terms that are also provided by a larger set of entities from a different source were removed.
- Sets of “equivalent” records (records providing the same set of “useful” terms) were identified. One representative record from each equivalence set was retained (with preference given to ChEBI records).
- Records were clustered based on shared terms. Large clusters were broken up by discarding records that were too generic.
- In each cluster, records that were redundant were removed.

Properties: All small molecule records have Synonym and CAS registry number properties. Additionally some records from ChEBI and PubChem also have Molecular WT and Formula properties.

Enzyme

All enzyme data were obtained from Expasy and MeSH.

Family

The family names were computed by scanning all Medline abstracts and extracting all sequences of words which satisfy the following three criteria:

- each word has occurred in the name/alias/description of some protein
- the first word has occurred as the first token of some protein
- the last word is either family/subfamily or the plural form of the last word in some protein

Approximately 110,000 unique potential family names were identified in this manner with some degree of manual curation. The E. coli database also includes approximately 299 operons as families. Family entities have no properties other than Synonym.

- **Process, Function and Complex** The process and function entities in the **GeneSpring GX** databases correspond exactly to the process and function sections of the GO hierarchy respectively. Gene Ontology (GO) is an hierarchical organization of process, function and cellular component terms that can assist in the annotation of proteins.

The complex entities in our database correspond to the GO IDs that are in section of the GO cellular component hierarchy rooted at GO:0043234 ("protein complex"). We do not have complex constituent information at present.

All properties of the entities are directly taken from the gene-ontology.obo file available at GO as of 2008.

Properties: All process Function and Complex properties have been deciphered from Gene Ontology

Entity Type	Properties
Process	Alias
	Description
	GO ID
	Name
	Note
Function	Alias
	Description
	GO ID
	Name
	Note
Complex	Alias
	Description
	EC Number
	GO Component
	GO ID
	Note

Table 25.3: Process, Function and Complex Properties

25.10.3 Relations

To extract relation information from literature, the GS Natural language processing algorithm uses a dictionary consisting of the above described entities. A relation represents molecular interactions between entities. A relation is characterized by a set of participating entities. Entities play several roles within each relation. The relations in the GS Pathway database are of the following types:

- Binding
- Expression

- Metabolism
- Promoter Binding
- Protein Modification
- Regulation
- Transport

Member (these relation are not included in the **GeneSpring GX** database, but can be provided upon request).

Each relation is characterized by two or more entities playing the following primary roles. We represent a basic relation with two participants in text format as $A \rightarrow B$. We refer to A as the “regulator” node and B as the “target” node. In addition, several entities may influence the relation of $(A \rightarrow B)$. These participants are called controllers. In the plain text format we represent them as $C \rightarrow (A \rightarrow B)$, where C is the controller. Some of these roles are defined with more specific names in the context of specific types of relations. Example: Controllers of a metabolism reaction are called Catalysts.

The “effect” attribute of a relation, reflects how the change (may be qualitative or quantitative) in A influences B. If an increase/decrease in A, is followed by an increase/decrease in B (same direction), then the effect of A on B is considered to be positive. If an increase in A, is followed by a decrease in B or vice versa, then the effect of A on B is considered to be negative. The various combinations and their text format representations are shown below: Change in A Change in B Effect Representation

We represent a regulation with unknown effect in text format as $A \rightarrow B$. Positive and negative effects are shown as $[A \text{ --+> } B]$, $[A \text{ --| } B]$ respectively.

The following table summarizes the various roles played by the participants of each type of relation, including their effects.

Explanation on the types of Relations:

- **Regulation**

“Regulation” is the most basic relation type. All relation types (except Member) in the above list can be viewed as more detailed versions of the regulation relations. We say that an entity, A, “regulates” another entity, B, if A has some influence on B. The participant entities in “regulation relation” are “regulator”, “target” and “modulator”. Each participant can be ascribed as playing a positive, negative or unknown effect.

The “mechanism” attribute indicates the way in which B was influenced by A. At present this attribute is always set to “None” for Regulation relations.

- **Binding**

Relation Type	Everyone	Left	Right	Controllers
Binding	Participant			Modulator (+) Modulator (-) Modulator
Member		Left	Right	Modulator (+) Modulator (-) Modulator
Metabolism		Reactant	Product	Catalyst (+) Catalyst (-) Catalyst
Regulation, Protein Modification, Promoter Binding, Expression, Transport		Regulator (+) Regulator (-) Regulator	Target (+) Target (-) Target	Modulator (+) Modulator (-) Modulator

Table 25.4: Participant Roles

Two entities, A and B, participate in a "Binding" relation, if they physically interact with each other. Since both A, B take part equally in the interaction, this is an undirected relation and there is no "regulator" and "target" node. Both entities are called "participants". The text representation takes the form $A - B$. If multiple entities, say A, B and C, bind together to form a complex, we can discard the line representation and depict it as A, B, C. There are no effects ascribed to the participants. However "modulators" which can exert either a "positive" or a "negative" or an "unknown" effect, can control binding relations.

Data from Two-hybrid screens, immunoprecipitation, and other laboratory interaction detection methods can best be represented as binding relations.

- **Promoter Binding**

If protein A binds to the 5' upstream (or promoter) region of gene B, we represent the relation as a "Promoter Binding" relation, $A \rightarrow B$. Note that this is a directed relation like **Regulation**, in which the target node is the gene entity. This relation type is characterized by "regulator", "target" and "modulator" participants exerting either positive, negative or unknown effects on the relation.

Transcription factor binding information can best be represented in the form of Promoter Binding relations.

- **Expression**

If A causes a quantitative change in the amount of protein/mRNA B, we represent the relation as an "Expression" relation. The participant roles are exactly the same as defined in Regulation.

- **Transport**

Relations involving the change of an entity because of it being transported between cellular compartments are represented by "Transport" relations. At present, we do not capture the mechanism of transport. The participant roles are exactly the same as defined in **Regulation**. The "target" entity is the molecule that is transported in the cell and the "regulator" entity regulates this transport.

- **Protein Modification**

Relations in which an entity A causes the post translational modification of a protein B are represented by "Protein Modification" relations of the form $A \rightarrow B$. The participant roles are exactly the same as defined in **Regulation**. The target entity is usually a protein. (However, we allow the target entity to be enzymes and family terms also because of the significant overlaps amongst these entity types). This type of relation has an additional attribute defined "mechanism" which explains the type of protein modification undergone by the target. The mechanisms included in **GeneSpring GX** are

- Phosphorylation
- De-Phosphorylation
- Glycosylation
- De-Glycosylation
- Ubiquitination
- De-Ubiquitination
- Sumoylation
- De-Sumoylation
- Methylation
- De-Methylation
- Acetylation
- De-Acetylation
- Myristoylation
- Palmitoylation
- ADP-ribosylation
- Geranylgeranylation
- Farnesylation
- Sulfation
- Unknown

At present this relation type is the only one which uses the mechanism attribute in a non-trivial manner. All other relation types have the mechanism attribute value set to "Unknown".

- **Member** This is used to store relations of the form "A protein is a member of family B", "A protein is a part of complex B". The text format is $A \rightarrow B$. These relations are not parsed from NLP and hence not included in the **GeneSpring GX** database currently. They can be provided to the customer upon request.
- **Metabolism** Relations involving the quantitative change of small molecule entities are represented as "Metabolism" relations. Example: "Enzyme E catalyzed the conversion of small molecule A to B". In this directed relation, the left participant A is the "reactant", and the right participant B is the "product". Reactants and products do not possess any effect. The controller of this reaction (relation) is enzyme E, which I called the catalyst for the relation.

25.10.4 Database statistics

The statistics for the **GeneSpring GX** pathway database are shown on this page.

Each organism database is characterized by its unique protein entities. Proteins and genes are not distinguished in the **GeneSpring GX** database.

Entities	Human	Mouse	Rat	Drosophila
Proteins/Genes	39964	63570	37676	22722

Table 25.5: Protein Entities in Pathway Database

Entities	C.elegans	Yeast	Arabidopsis	E.coli
Proteins/Genes	21052	6200	33263	42564 (all species)

Table 25.6: Protein Entities in Pathway Database

The other entities are common to all organisms. We have an additional 299 families in E.coli representing the "operons" as protein families.

Entities	Number
Small Molecules	55376
Enzymes	4750
Function	8260
Complex	902
Process	15005
Family	113196 (+299 in E.coli)

Table 25.7: Other Entities in Pathway database

Relations that are applicable to all organism are stored in the "Generic" database. This database contains:

Relation	Count
Binding	21448
Expression	388
Metabolism	20142
Promoter Binding	13
Protein Modification	0
Regulation	81400
Transport	3125

Table 25.8: Total Number of Relation classified as "Generic"

The number of relations in each organism database are given in the table [25.10.4](#).

We report the total number of unique relations that were obtained from each data source (NLP, IntAct

Relations	Human	Mouse	Rat	Drosophila
Binding	132940	54632	40150	34749
Expression	106972	57576	32054	3140
Member	0	0	0	0
Metabolism	107619	49179	61319	4355
Promoter Binding	5724	2759	927	505
Protein modification	30150	15456	10262	1570
Regulation	974925	472312	580991	36537
Transport	57140	22811	41593	1234

Table 25.9: Total Number of Relation in Pathway database

Relations	C.elegans	Yeast	Arabidopsis	E.coli
Binding	11254	49248	5754	23067
Expression	844	2496	1277	11504
Member	0	0	0	0
Metabolism	5275	3955	2015	10599
Promoter Binding	17	4145	79	873
Protein modification	838	1206	611	852
Regulation	23962	32889	13790	59282
Transport	932	1053	392	2199

Table 25.10: Total Number of Relation in Pathway database

and other sources) for each organism. Clearly the NLP algorithm captured the maximum number of relations for all organisms. The experimentally reported physical interactions data was parsed from IntAct (www.ebi.ac.uk/intact), (<http://www.ebi.ac.uk/intact>). Our in house analysis (not reported here) shows that IntAct includes data from other databases like BIND and MINT. The actual number of relations parsed from IntAct were larger than those reported here. These numbers represent only the relations whose sole reference is in IntAct. Note that IntAct database gave a substantial number of unique interactions for Drosophila and Yeast. For all organisms, several relations had associate references derived from NLP as well as IntAct. We have also enriched the Yeast database with interactions found in the SGD. Future efforts will try to obtain data from additional sources, for all model organisms.

Source	Human	Mouse	Rat	Drosophila
NLP	116935	673859	766296	58082
IntAct (unique)	16025	866	1000	24006

Table 25.11: Relations from each Data source

Source	C.elegans	Yeast	Arabidopsis	E.coli
NLP	38744	46518	21611	94670
IntAct (unique)	4378	31847	2307	13076
Other sources	0	16627	0	0

Table 25.12: Relations from each Data source

25.10.5 Overview of Natural Language Processing (NLP)

A majority of relations in the **GeneSpring GX** Pathway database are derived using a Natural Language processing Algorithm that runs on published Medline abstracts until January 2008. NLP is based on a "deep parsing" method and driven by an elaborate sentence grammar, that maximizes accuracy and has control over different aspects of a sentence without compromising recall. The NLP system operates on a sentence-by-sentence manner and extracts only those relations that are completely within a sentence. The four main phases - entity recognition, syntax analysis, semantic analysis, semantic inferencing - are described in greater details below.

- **Entity recognition:** Entities occurring in the sentence are tagged by consulting entity dictionaries. Since there are variations in how terms appear in the literature (hyphens, Greek letter), we use a tokenization procedure, which breaks every term into tokens, identifies matching tokens and strings back collection of matching tokens into terms. Some of the cases handled by entity recognition are:
 - alternate ways of writing the same name: NFkappa-B, NF Kappa B
 - local abbreviation expansion: abbreviations defined early in the abstract are expanded in the rest of the abstract before entity recognition. Eg. caspase 1, 3 is recognized as caspase 1, caspase 3.

Only sentences in which at least two entities have been recognized are passed on to the next stage.

- **Syntax Analysis:** Using a set of rules, the syntactic tree structure of the sentence is derived using context free grammar rules for English. The syntactic structure breaks up the sentence into its underlying linguistic constituents like noun, verb, and adverbial/propositional clauses and phrases. It also captures the functional roles of different parts of the sentence like the subject, object, subject modifier, object modifier and predicate modifier. Before entity recognized sentences are parsed for syntax tree, a lexical analysis step refers each word to a dictionary of English words derived from UMLS Lexicon. This would handle variations caused by tense, number, person etc.

Several different syntactical structures of sentences could carry the same semantic context:

- A regulates B
- B is regulated by A
- A plays a role in the recognition of B
- A belongs to a family of B regulators
- B activity was found to be regulated by the addition of A

The next two stages: Semantic analysis and inferencing recognize sentences with different structures carrying the same meaning.

- **Semantic Analysis:** This phase converts all syntax trees into semantic trees using a semantic dictionary that maps all words of interest to "semantic" concepts. For eg. The word "modulate" would map to "regulate". The semantic tree would then need to identify what regulates what using the sentence structure imposed by the syntax tree. The relationships captured by the semantic tree are only the direct ones. These are only a fraction of the total relations. Indirect relations can be captured by making inferences across multiple relationships present.

- **Inferencing:** An example of a sentence that requires inferencing is as follows "treatment by A causes an increase in the amount of B". The semantic tree would yield the following concepts: cause, treatment, increase, amount. To conclude that A regulates B, we would need to make inferences across these four semantic concepts, using agents in one relationship to fill the missing holes in other relationships. The process is driven by domain specific inference rules which make several passes through the semantic tree trying to unify the various concepts present in the semantic tree and inferring new concepts. We can augment the semantic network adding new semantic nodes that would represent the inferred concepts. We extract relations by searching the resulting semantic network for relation nodes with all its arguments filled.

A reference consisting of the following fields is created for each parsed relation:

- PMID (PubMed ID)
- actual sentence
- year of publication
- journal name
- reference score
- source

A signature is created for each relation depending upon the participants, their roles and mechanisms. References to the relation are added based upon the signature derived from the reference sentence.

25.11 Update Pathway Interactions

GeneSpring GX Pathway database will be updated with relations derived from the latest PubMed abstracts on a regular basis. These additional relations can be updated from time-to-time from the user interface from the *Annotations*→*Update Pathway Interactions*. Two different options are available:

- From Agilent server- using the web
- From files- using a file based update

See figure [25.49](#)

25.12 Working with the Pathway Interactions Server

Note that in Workgroup mode, the *Annotations*→*Update Pathway Interactions* menu item is disabled. Contact your Systems Administrator to update the Pathway Interactions Server with the appropriate organism databases. The Pathway Interactions Server is a centralized server which stores organism databases.

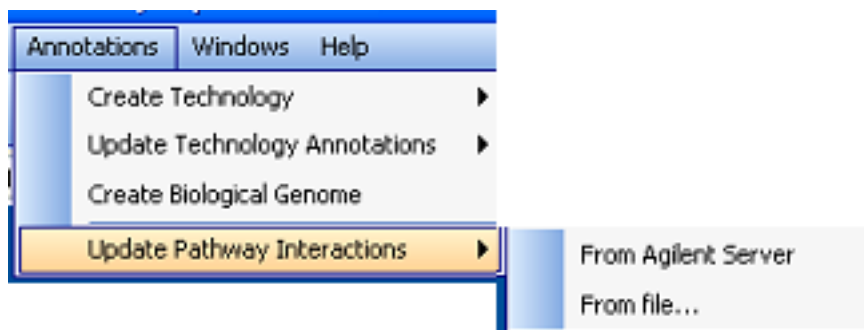


Figure 25.49: Update Pathway Interactions

See *Tools* → *Options* → *Pathway* → *Pathway Server Settings* to point your **GeneSpring GX** client to this server.

Only an administrator can update organism databases via a web page. In addition, only an administrator can create a new organism; this needs to be done via the **GeneSpring GX** client using *Annotations* → *Create Pathway Organism*. Once this is done, all **GeneSpring GX** clients can use all the pathway analysis features listed above. Note that entities, relationships and their properties in the Pathway Interaction Server do not have any user access controls; all users can query these and write back new entities and relationships. Pathways derived from these entities and relationships however are governed by the usual Workgroup user access rules. A key feature of the Pathway Interactions Server is the speed with which pathway analysis queries are processed compared to the desktop version.

25.13 Troubleshooting

For the desktop version, the key areas of trouble are the following:

- You do not have a pathways license. Contact **GeneSpring GX** support (informatics_support@agilent.com) to get a pathways license.
- You have not downloaded the pathways database for your organism of interest. Go to *Annotations* → *Update Pathway Interactions*.
- Your MySQL database has not started up. Check the task manager for a running `mysqld-nt.exe` process. If you do not see this process, try restarting, and contact **GeneSpring GX** support if needed.
- You get a connection error after **GeneSpring GX** has been on but idle for several hours. You will need to restart **GeneSpring GX**.

For the Workgroup version, the key areas of trouble are the following:

- You do not have a pathways license. Ask your administrator to contact **GeneSpring GX** support (informatics.support@agilent.com) to get a pathways license.
- You can't connect to the pathway server; check that the pathway server information is correct in *Tools*→*Options*→*Pathway Server Settings*.
- Your pathways interaction server is not updated with the database for the organism of your interest. Ask your systems administrator to update the server or create a new organism on the server as the case may be.

And whenever you contact **GeneSpring GX** support, please remember to provide the `stderr.log` and `stdout.log` files along with all files in the `bin/launcher/lib/logs` subfolder; these files are rewritten on each restart so remember to capture these files the moment you see a problem.

Chapter 26

Copy Number Analysis

26.1 Introduction

This chapter introduces Copy Number, Allele Specific Copy Number, Parent Specific Copy Number, LOH, Log Ratio and Common Genomic Variant Regions, which are new features in **GeneSpring GX 11.0**.

It was generally thought that genes were almost always present in two copies in a genome, resulting in a 'Copy Number' of 2. But recent studies have proved that the human genome shows extensive Copy Number variation (CNV). In humans, CNVs encompass more DNA than single nucleotide polymorphisms (SNPs). Like other types of genetic variation, some CNVs have been associated with susceptibility or resistance to disease (Pathogenic). CNV analysis includes detection of regions whose Copy Number varies in comparison to a reference genome.

For each SNP allele, denoted by the letters A and B, the individual Copy Numbers of the two alleles are referred to as Allele Specific Copy Number (ASCN); their sum adds up to the total Copy Number. This can be assigned to the genome from the total Copy Number, signal intensities of the alleles and the genotype call. Another way to split up the total Copy Number into components is based on Copy Numbers in individual chromosomes inherited from each parent; these are called Parent Specific Copy Number (PSCN). This information is derived from the total Copy Number and the Allele Specific Copy Number. Allele Specific Copy Number and the Parent Specific Copy Number help in tracing the chromosomal amplifications/deletions at the allele level. Loss of Heterozygosity (LOH) refers to change from a state of heterozygosity in a normal genome to a homozygous state in a tumor genome. LOH can result from copy-loss events such as hemizygous deletions (Copy Number dependent), and also from copyneutral events such as chromosomal duplications (Copy Number independent).

Log Ratio is a simple ratio of the summarized intensities of the sample versus reference or normal as the case may be, in log (to the base 2) scale.

All the above computations follow either the **Against Reference** method or the **Paired Normal**

method. The former method involves comparing specified arrays against a reference file created from a set of reference individuals. The latter compares each specified array against a corresponding array obtained from a normal tissue sample of the same individual. The sections below explain in detail the various steps of Copy Number analysis which includes computation of Copy Number, ASCN, PSCN, LOH score and Log Ratios.

Determining **Common Genomic Variant Regions** is another feature in **GeneSpring GX 11.0**, which identifies regions of the genome that are significantly amplified or deleted across a set of samples. **GeneSpring GX 11.0** follows the procedure described in this reference on Genomic Identification of Significant Targets in Cancer (GISTIC) [23]. For complete details on implementation, see section [Common Genomic Variant Regions](#).

GeneSpring GX supports genotyping analysis for Affymetrix and Illumina. In case of Illumina, outputs from GenomeStudio are supported.

26.1.1 Terminology in Copy Number analysis

Commonly found terms with respect to Copy Number analysis are explained in Table 26.1

26.2 Technologies supported by GeneSpring GX 11.0

GeneSpring GX 11.0 supports Affymetrix CEL files and Illumina output files from GenomeStudio (extracted using a plug-in available for **GeneSpring GX 11.0**) for Copy Number analysis and the computation follows slightly different workflows for the different technologies listed below.

- **Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set**
- **Affymetrix Human Mapping 100K Set**
- [Illumina Genotyping output files](#)

The primary aim of Copy Number analysis is to segment the genome according to the Copy Number. For Affymetrix technology, the following steps broadly constitute the Copy Number analysis.

26.2.1 Experiment Creation

GeneSpring GX calls Affymetrix Power Tools (APT) (<http://www.affymetrix.com/support/developer/powertools/changelog/apt-probeset-summarize.html>) for carrying out summarization of the CEL files

Table 26.1: Terminology in Copy Number Analysis

Copy Number	Number of copies of a particular gene within the genome; Common Copy Number is 2, but human genome has been found to show extensive variation in Copy Number.
Copy Number Variants	A variant region is classified as Copy Number Variant (CNV) only if it is of at least 1Kb in size. CNVs can be Copy Number gains (insertions and duplications) or losses (deletions) relative to the reference genome. CNVs can be de novo or inherited. Most variations are benign; assigning a variation to pathogenicity needs a genome wide association study
Copy Number Variants Region	Copy Number Variants Regions (CNVRs) represent merging of several overlapping, but independently ascertained CNV segments. CNVRs can have large number of potential Copy Number variants within it.
Copy Number Polymorphism	Copy Number Polymorphism (CNP) are subset of CNVs that segment at a frequency of one percent in the population. Large-scale Copy Number polymorphisms are common and widely distributed in the human genome.
Copy Number Confidence	CBS assigns a confidence value for the segments; For each segment with a mean log ratio value, a T test is done against the mean '0' (which indicates Copy Number 2) and a p value is obtained. Negative logarithm to the base 10 of this p value is reported as confidence.
Log Ratio	Log Ratio is the ratio of the summarized intensities of the sample versus the reference or the normal, as the case may be, on log to the base 2 scale.
Mean Log Ratio	For each Copy Number segment, CBS gives the 'Mean Log Ratio'. This value can be viewed in genome browser after running Copy Number analysis.
Copy Neutral LOH	Loss in Heterozygosity (LOH) that results from events which do not involve change in Copy Number; for instance, chromosomal duplication. The LOH score is between 0 and 1; Score over 0.5 is considered to be LOH by default; but the value is configurable.
Homozygous, Heterozygous and Hemizygous condition	If both alleles in a gene are same, it is homozygous; if they are different, it is heterozygous; if one is missing, it is hemizygous
Against Reference Analysis	Compares specified arrays against a reference file created from a set of reference individuals
Paired Normal	Compares each specified array against a corresponding array obtained from a normal tissue sample of the same individual
Allele Specific Copy Number	Assigning allele calls to the SNP using Fawkes Algorithm
Parent specific Copy Number	GeneSpring GX 11.0 assigns a split in the total Copy Number to infer the parent specific Copy Number, from the Allele Specific Copy Number. Refer to section PSCN for details.
p value in assigning Allele Specific Copy Numbers	It is the p value of the Fawkes result; check p values of fawkes result for details. The lower the value is, the better the confidence in prediction is.
GISTIC	Identifies common genomic variant regions ; commonly called Genomic Identification of Significant Targets in Cancer; see Common Genomic Variant Regions .

using the PLIER algorithm. See [Probe Summarization Algorithms](#) for details on PLIER algorithm. Some of the PLIER parameters are :

- For Quantile Normalization, default of 1 percent or 50,000, whichever is higher, is used.
- Sets all chips median to 1000.
- Simplified Expression Analysis (SEA) optimization method is used.
- Other parameters are set to default in APT.

For Affymetrix Mapping 100k array set, [BRLMM](#) algorithm is also run during experiment creation and the outputs are stored. Figure [26.1](#) explains experiment creation for all Affymetrix technologies.

26.2.2 Reference

During experiment creation, there is an option to choose the standard reference prepackaged within the tool, or to use a custom reference created by the user.

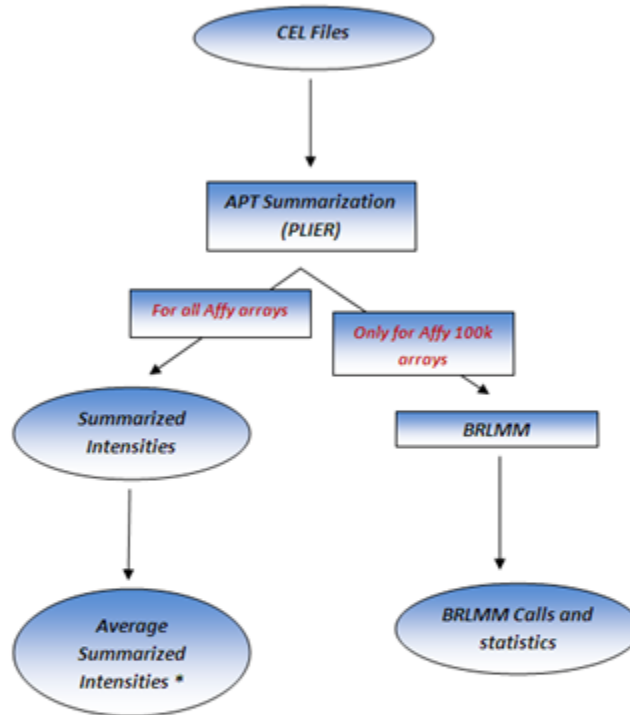
GeneSpring GX comes prepackaged with a reference created using the Phase II 270 HapMap samples, [\[27\]](#), the second generation human haplotype map from The International HapMap Consortium. According to the reference cited above, this Phase II HapMap characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25 to 35 percent of common SNP variation in the populations surveyed.

If the user has reasons to believe that the 270 HapMap samples of the standard reference will not be a good reference for his data, and if he has access to data that is better suited for the comparison, then he can create a custom reference with those samples. Note that the samples used for custom reference should belong to one of the standard technologies supported by the product. The reference creation procedure is the same for both standard and custom references and is shown in this figure [26.2](#) for the Affymetrix Genome-Wide Human SNP Array 6.0, the Genome-wide Human SNP array 5.0, and the Human Mapping 500K Array Set.

Handling sex chromosomes during reference creation: The reference could be a mix of female and male samples and female samples will have a higher signal for X chromosome compared to male samples. To take care of this, signals for chromosome X from male samples are scaled by a factor of 1.4 for for Affymetrix Mapping 100k array set (Hind, Xba and paired). With this scaling, the reference effectively becomes an 'all female' sample set.

For all the other arrays (250k, 500k, V5, V6), the average signals are replaced by the median of the copy number 2 clusters obtained from Birdseed algorithm. This effectively accounts for the scaling and

Flow Chart showing Experiment Creation for Affymetrix Technologies



• For SNP probe, it is the average of the two allele intensities. For CN probes, it is the same as Summarized intensities.

Figure 26.1: Experiment Creation for Affy CEL files

weighting required for male samples. Birdseed algorithm does not run on Affymetrix Mapping 100k array set and hence the explicit scaling for this array set.

For the Affymetrix Mapping 100k array set, BRLMM is run in place of Birdseed (and Canary) and the procedure is shown in Figure 26.3.

26.2.3 Copy Number Analysis

Once an experiment has been created, Copy Number analysis can be performed. This step uses the following information:

- The Summarized dataset.
- The chosen Reference.
- Interpretation for deducing if it is 'Against Reference' analysis or 'Paired Normal' analysis. See section [Significance of Interpretation in Copy Number analysis](#) for details.

From the above inputs, Log Ratios, Copy Numbers, LOH scores, Allele Specific Copy Numbers and Parent Specific Copy Numbers are calculated.

Log ratios:

- **Against Reference:** The log ratio is the ratio of sample probeset intensity to reference probeset intensity (i.e., the average of the probeset intensities over all the arrays in the reference), transformed by logarithm to base 2. For SNP probes, the intensity used is the average of the individual A and B intensities. By way of normalization, the median of the sample probeset intensities is scaled to the median of the reference probeset intensities before log ratios are computed.

- **Paired Normal:** For disease/tumor samples, the log ratio is the ratio of tumor probeset intensity to the normal probeset intensity, transformed by logarithm to base 2. For SNP probes, the intensity used is the average of the A and B intensities. Log ratios are then median shifted to 0 for each disease/tumor sample. For normal samples, the log ratio is generated using the Against Reference computation above.

Copy Number: Log Ratios are fed to a CBS segmentation routine for segmentation into regions of roughly equal log ratios. For each resulting segment, a confidence value ($-\log_{10}(\text{pvalue})$) is computed; higher confidences indicate segments in which the mean log ratios are well away from 0 and the variance in the log ratios is small. For each segment, we also compute a copy number value, using a calibration method that associates the mean log ratios in that segment with a copy number value. The copy number values output are continuous values up to the first decimal digit and range from 0 to 4.

See sections [Circular Binary Segmentation algorithm \(CBS\)](#) for segmenting the genome on the basis of Copy Number. CBS outputs segment start and end points along with the Mean Log Ratios for

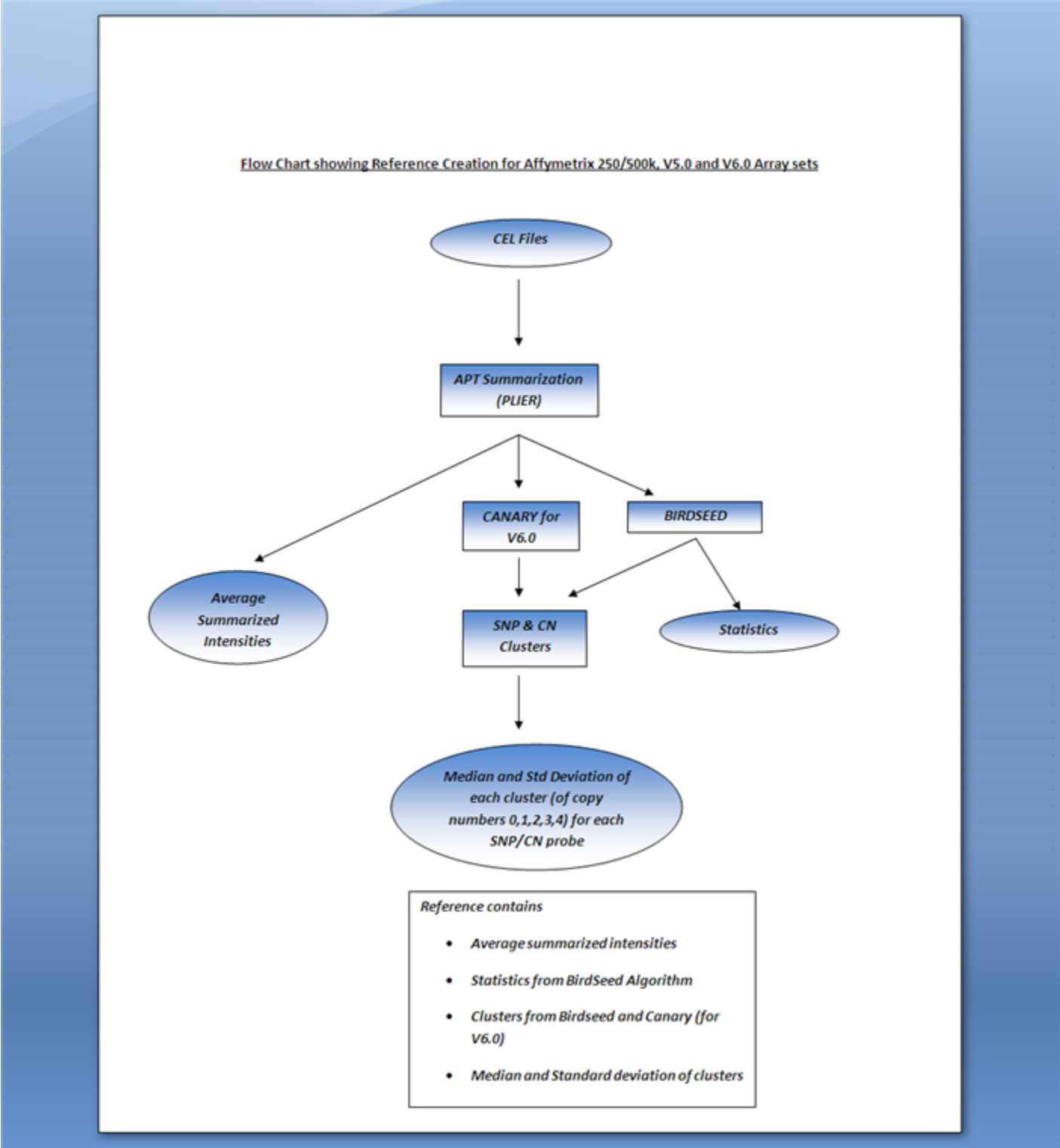


Figure 26.2: Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set - Reference Creation

Flow Chart showing Reference Creation for Affymetrix 50/100k Array set

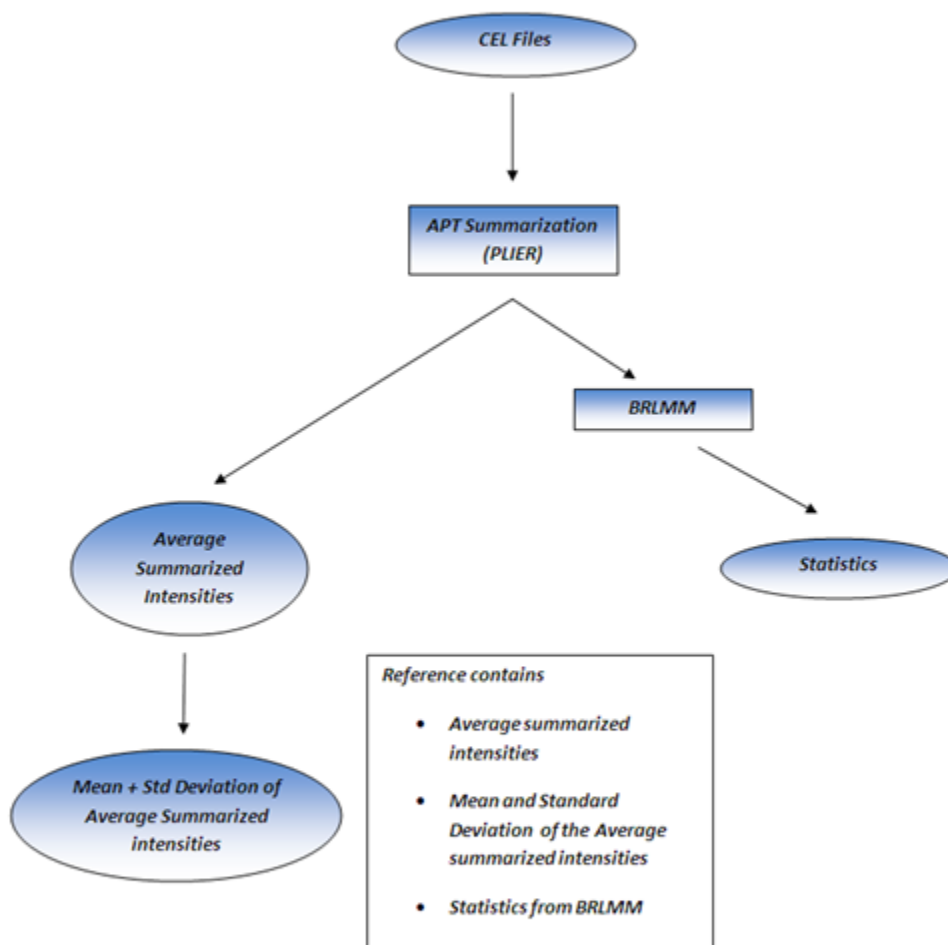


Figure 26.3: Reference Creation for Affy 100K array set

each segment. Copy Numbers are assigned to segments based on a mapping explained in section [Post Processing to assign Copy Numbers to segments from CBS](#).

Allele Specific Copy Number: For SNP probesets, the Total Copy Number is split into its A and B allele specific components using the Fawkes algorithm. This computation is not performed for the Affymetrix Mapping 100K technology. The total copy number is rounded off to the nearest integer before being fed to Fawkes and the ASCN's output by Fawkes are always integral.

Parent Specific Copy Numbers (PSCN): Total Copy Numbers are split into Parent Specific Copy Numbers, i.e., individual copy numbers for each of the 2 chromosomes. The intuition behind this computation is as follows. Consider a copy number segment with say copy number 3, and consider all SNPs in this segment. If all of these have ASCN A=2,B=1, we can conclude that one of the chromosomes has 2 copies and the other has 1 copy. On the other hand, if all of these have ASCN A=3,B=0, then it is not possible to unambiguously determine the split into PSCNs (all 3 A's could be on one chromosome or 2 A's on one and 1 A on the other). So PSCN inference is based on heterozygous SNPs and nearby homozygous SNPs inherit PSCN's inferred from heterozygous SNPs. Long stretches of homozygous SNPs lead to inconclusive PSCNs and are given missing values. The stretch to be called as homozygous can be set from the configuration options from the 'Tool' menu (default value is 5). For example, if there are 3 probe-sets with 2,0 ASCN split and the neighbouring ASCN values are 1,1 then the entire region is assigned a 'diff' value of 1 (even for the 2,0 region where the diff is expected to be 2).

Inputs are the total Copy Number values for segments and the Allele Specific Copy Number. A 'diff' which is the difference between the ASCN values is assigned; the 'diff' values are a measure of 'Allelic Imbalance'. Along with the 'diff' value, the minimum (min) and maximum (max) values of the ASCN split can be viewed in the genome browser.

- For Copy Number 2, the PSCN possibilities are 2,0 (homozygous) and 1,1. For this segment of total Copy Number 2, **GeneSpring GX** detects and displays segments with predominantly 1,1 split by ignoring the homozygous points in between.
- For total Copy Number 3, the possibilities are 3,0 and 2,1. Homozygous regions (3,0) are ignored and the regions with predominantly 2,1 split are displayed with minimum and maximum.
- For total Copy Number 4, the possibilities are 4,0 and 3,1 and 2,2. Homozygous regions (4,0) are ignored; in the remaining regions, the homozygous points are again removed and CBS is run on this. The homozygous points that are ignored should be at least 5 SNPs long, by default; this value is configurable. On the segments that are returned by CBS (which include blanks for homozygous points), we get the median and populate the min and max using this value.

LOH Scores HMM Implementation computes LOH scores. The inputs are BRLMM calls in case of Human Mapping 100K Set and Fawkes outputs in case of Genome-Wide Human SNP Array 6.0, Genome-Wide Human SNP Array 5.0 and Human Mapping 500K Array Set.

The possible Fawkes states and their mappings are presented in the table [26.2](#)

- **Against Reference:** For computing LOH scores for all samples, HMM (Against Reference version) is run using the genotype calls of the samples generated using Fawkes and some statistics about the the standard reference genotype calls dataset saved in the standard reference packaged with the technology.

Fawkes State	Mapping for LOH
A0B0	No Call
A0B1	BB
A1B0	AA
A0B2	BB
A2B0	AA
A1B1	AB
A0B3	BB
A1B2	AB
A2B1	AB
A3B0	AA
A0B4	BB
A1B3	AB
A2B2	AB
A3B1	AB
A4B0	AA

Table 26.2: Mapping Fawkes state to LOH

In human genome, the normal level of homozygosity is around 70 %. When doing an 'Against Reference' analysis, any region with more than 70 % homozygosity can be safely assumed to be showing LOH. This combined with some meaningful filter conditions under '[Identify Copy Neutral LOH](#)' can reduce false positives and remove small pockets of high homozygosity while picking up LOH regions.

- **Paired Normal:**

- In Fawkes, the calls for Normal samples are set to 2 before sending to LOH computation.
- The Strand LOH HMM (Paired Normal version) is run to calculate the LOH scores for the tumor samples
- The Strand LOH HMM (Against Reference version) is run to calculate the LOH scores for the normal samples

26.2.4 Special mention for Affymetrix Mapping 100k Array

- BRLMM needs at least 6 distinct CEL file pairs or at least 6 distinct CEL files. Hence for Affymetrix Mapping 100k Array set, if the number of file pairs (or files) is fewer than 6, experiment creation is aborted.
- Allele Specific Copy Number and hence Parent Specific Copy Number cannot be generated for Affymetrix Mapping 100k Array.
- Post CBS, SNP/CN clusters from reference are used for the mapping to assign Copy Numbers to segment. For Affymetrix Mapping 100k array set, since there are no CN clusters (only BRLMM runs on this array), clusters generated from Affymetrix Genome-Wide Human SNP Array 6.0 technology of the reference are used here.

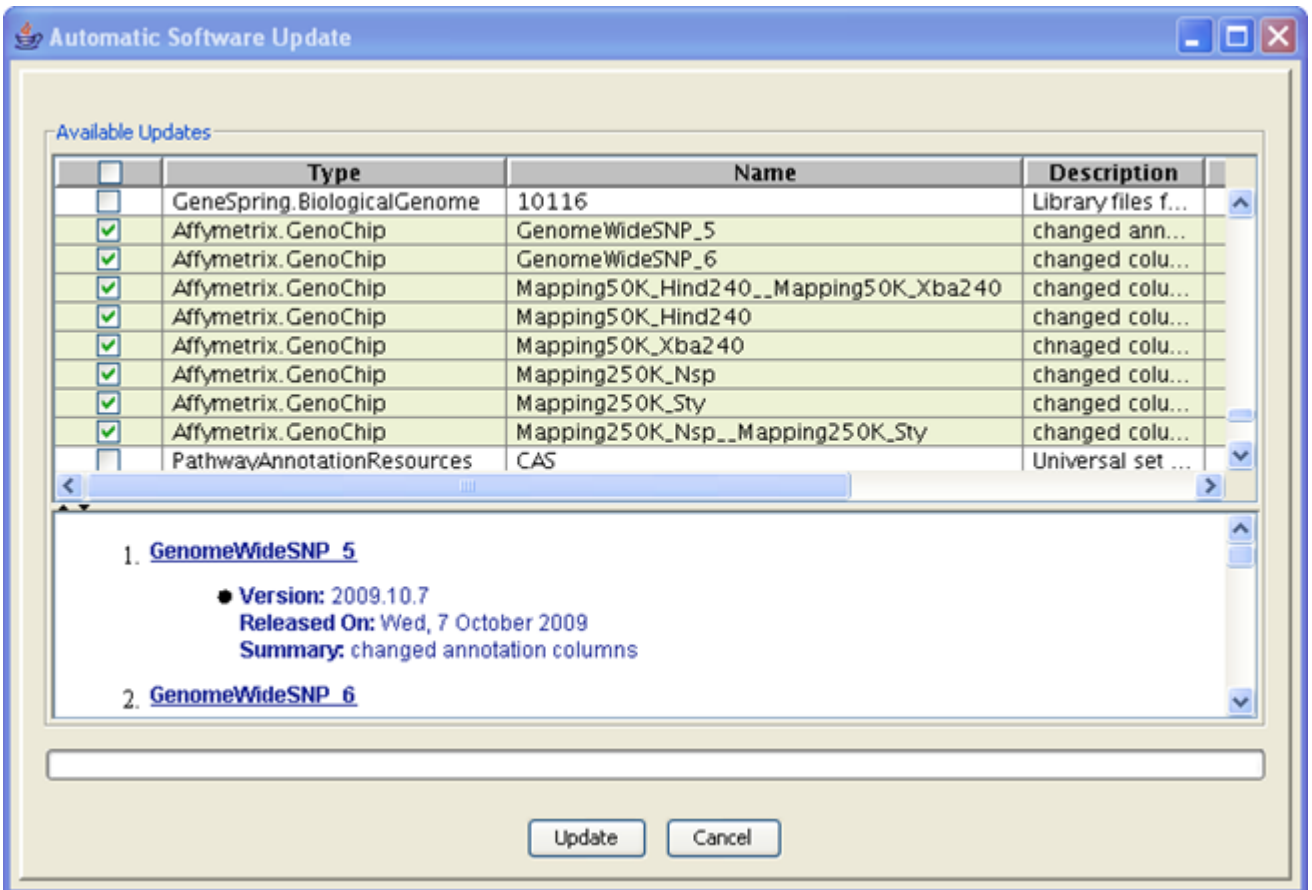


Figure 26.4: Create Technology for Copy Number Analysis - Affymetrix technology

The workflow and utilities common across all the technologies are described in the sections [Workflow description for Affymetrix Technology](#) and [Workflow description for Illumina](#). The algorithms used for various computations are explained in the section [Copy Number Algorithms](#).

26.3 Workflow description for Affymetrix files

26.3.1 Create Technology

Before starting Copy Number analysis, ensure that the required technology is downloaded. Go to the menu **Annotations** → **Create Technology** → **From Agilent Server**. From the list of available updates, look for those with the type 'Affymetrix GenoChip' along with the name of the exact technology. Technology downloads are available for Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set (individual and paired arrays) and Affymetrix Human Mapping 100K Set (individual and paired arrays). The download size is listed along side and the tool gives an estimate of the actual space required before starting the download. See [Figure 26.4](#).

Note that the appropriate technology should have been created before creating custom reference.

26.3.2 Creating a Copy Number experiment

To start the Copy Number analysis, an experiment has to be created with the Affymetrix CEL files in **GeneSpring GX**. This can be done by creating a new experiment within any existing project or by creating a new experiment in a new project.

1. Creating new Experiment in existing project:

Go to menu *Project* → *New Experiment*. Give a name for the experiment and choose the experiment type as 'Affymetrix Copy Number'. Note that for Copy Number analysis, only the advanced workflow is supported. The User can define experiment notes if any and click 'ok'. More details on creating new experiment are described below.

2. Creating new experiment in new project:

Alternately, close the existing project and go to menu *Project* → *New Project*. A window comes up prompting user to define a name for the project, as well as any notes. On clicking 'OK' here, an *Experiment Selection dialog* window appears with two options:

- **Create new experiment:** This allows the user to create a new experiment. (steps described below).
- **Open existing experiment:** This allows the user to use existing experiments from previous projects for further analysis.

Clicking on **Create new experiment** opens up a 'New Experiment' dialog in which 'Experiment name' can be assigned. Under experiment type, choose *Affymetrix Copy Number* from the drop down menu. Note that in Copy Number analysis, only *Advanced workflow analysis* is supported. Additional notes can be defined by the user in this window. Click 'OK' to begin the experiment creation.

A 3 step wizard opens up for creating a new Affymetrix Copy Number Experiment.

Step 1 of 3 - Load Data: This step allows loading of files/samples. *Choose Files* brings up a file chooser to choose 'CEL' files. *Choose Samples* brings up the sample search wizard. Refer [Search](#) for details on how to perform search. Figure 26.5 shows 'Load Data' step.

Once done, click *Next* to proceed.

Step 2 of 3 - Pair CEL files: This step provides the file pairing interface, and is shown only when CEL files from Hind and Xba arrays are input together. See Figure 26.6

This step is skipped if CEL files from *Genome wide Human SNP array 6.0* or *Genome wide Human SNP array 5.0* are used; it is also skipped when Hind and Xba files are separately input.

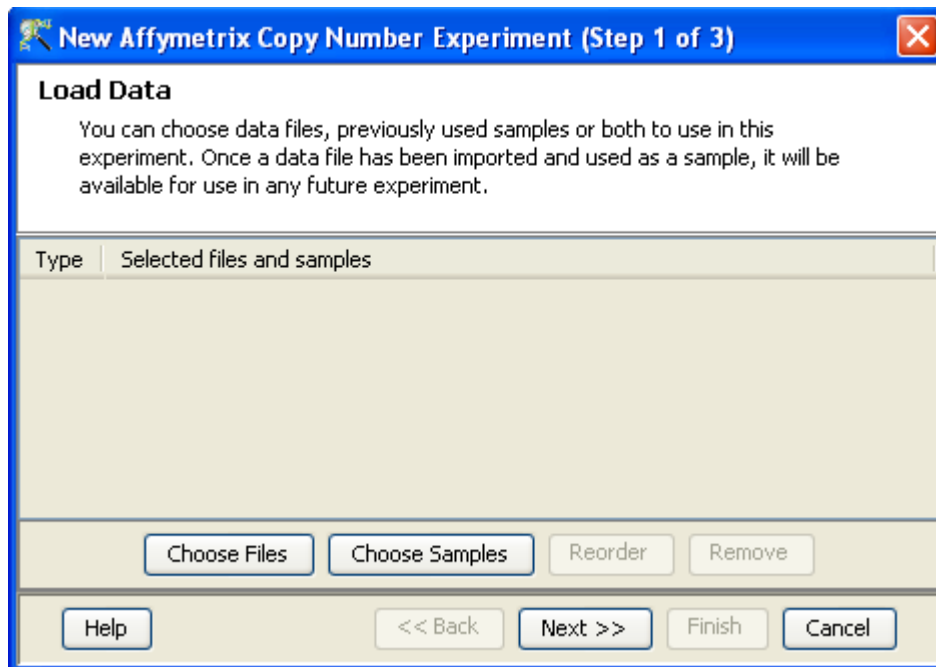


Figure 26.5: Step 1: Load Data

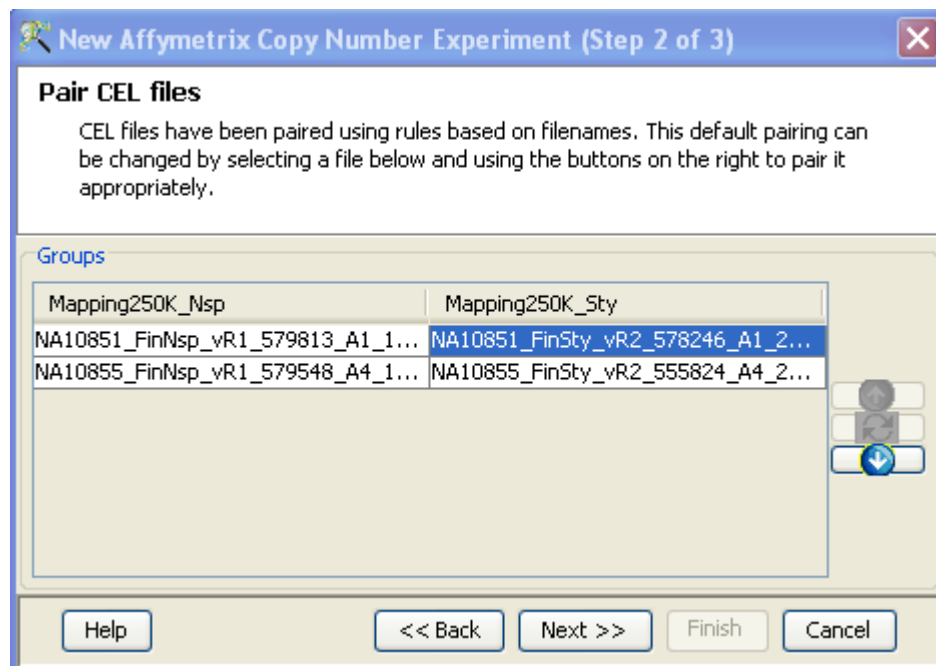


Figure 26.6: Step 2: Pair CEL files

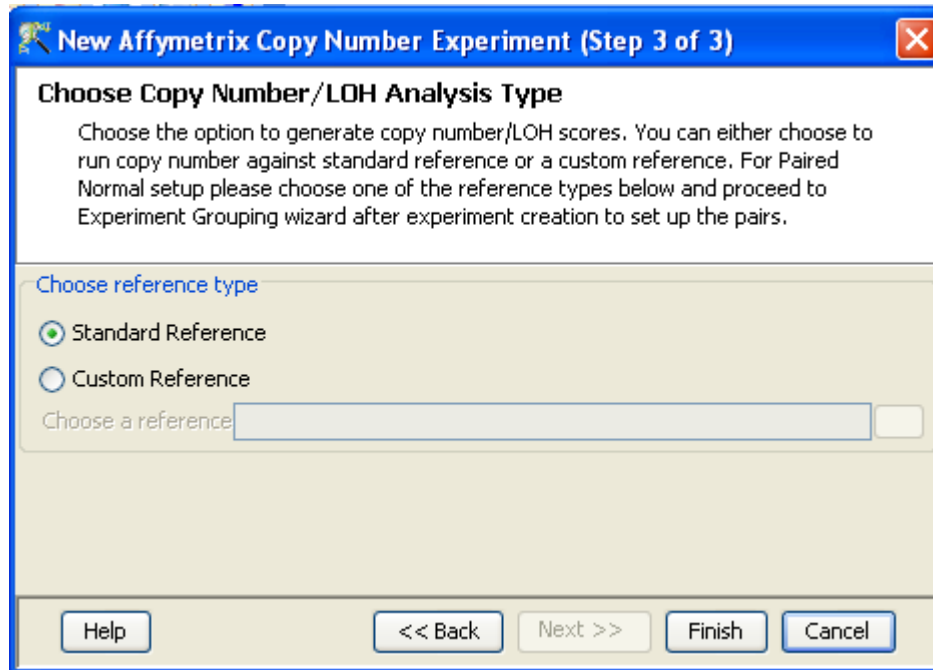


Figure 26.7: Step 3: Choose Copy Number/LOH Analysis Type

Both the 100K arrays and the 500K arrays comprise of two arrays of half the size each (the 100K arrays have Xba and Hind arrays of size 50K each and the 500K arrays have NSP and STY arrays of size 250K each). **GeneSpring GX** will attempt to automatically pair up the arrays based on naming rules. You can change the pairing or the order, by clicking on the appropriate cell and clicking on appropriate arrows.

NOTE: The order of the columns in the dataset will be the same as the order in which they occur in the selection interface. If you want the columns in the dataset to be in any specific order, you should order them here appropriately.

Step 3 of 3 - Choose Copy Number / LOH Analysis Type: This step allows the user to define options for carrying out Copy Number computation. **Choose Reference Type** option gives a choice to select pairing against standard reference or custom reference. See section [Technologies supported in GeneSpring GX](#) for details.

Figure 26.7 shows step 3.

See section [Custom Reference Creation](#) for details on creating custom reference. Once custom references are created using the procedure explained here, user can pick them for using here. Choose the required one and click *Finish* to exit the experiment creation wizard.

Technical details on what happens during experiment creation are explained in section [Experiment Creation](#).

26.3.3 Experiment Setup

Quick Start Guide: Clicking on this link will take you to the appropriate chapter in the on-line manual of **GeneSpring GX** .

Experiment Grouping: Experiment parameters defines the grouping or the replicate structure of the experiment. For details refer to the section on [Experiment Grouping](#).

Create Interpretation: An interpretation specifies how the samples should be grouped into experimental conditions both for visualization purposes and for analysis. For instance, if you have different types of tumours and you want view these in batches in the Genome Browser, you can create appropriate interpretations. For details refer to the section on [Create Interpretation](#). See section [Significance of Interpretation in Copy Number analysis](#) on how to define the interpretation to run Paired normal analysis.

26.3.4 Quality Control

Quality control on samples

Quality Control or the Sample QC lets the user decide which samples are ambiguous and which are passing the quality criteria. Based upon the QC results, the unreliable samples can be removed from the analysis. The QC view shows three tiled windows:

- 3D PCA scores
- 2D PCA scores, Experiment Grouping and Eigen values
- Legend

The *Add/Remove Samples* button allows the user to remove the unsatisfactory samples and to add the samples back if required. Note that the summarization is not re-done upon sample removal.

The QC view can be launched by using the workflow handle **Quality Control** → **Quality Control on Samples**. See Figure [26.8](#).

Batch Effect Correction

Batch effect is an example of systematic error which arises when different samples are run under slightly different conditions, for instance on different days. This step identifies and corrects markers that show consistent signal within batches but large variations between batches.

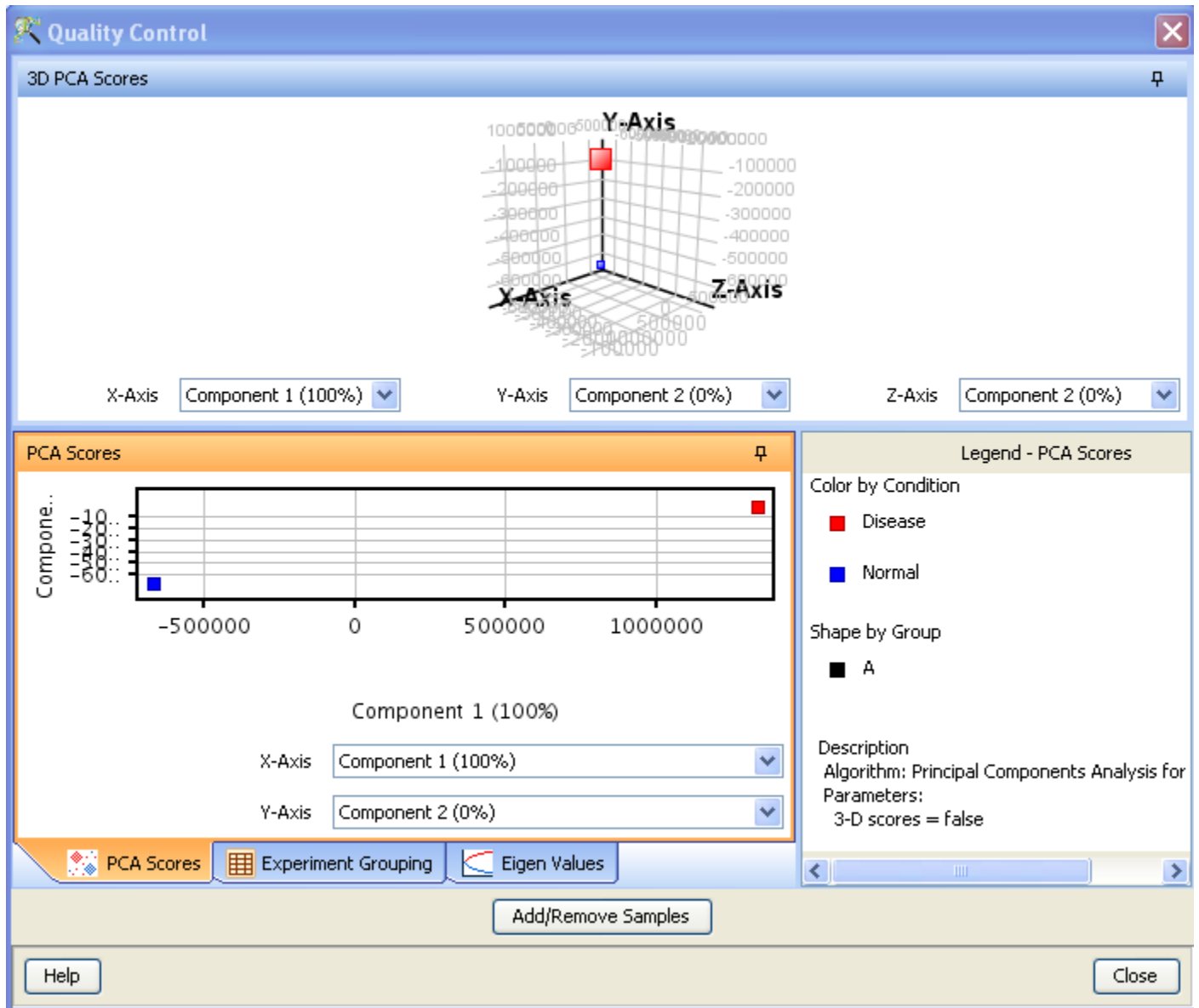


Figure 26.8: QC views for Copy Number Experiment

Batches are defined by conditions in the selected interpretation. Batches below a certain threshold number of samples are ignored in this procedure. Each remaining batch is T-tested against a pool of all the remaining batches. An entity is flagged for correction if it obtains a p-value below the specified threshold in the specified fraction of batches.

Process:

1. For every marker that has been marked for correction in a batch, b1, calculate the mean of the values for that marker in all the samples in all the batches that are NOT marked for correction. Let us call this ref-mean.
2. Calculate the mean of the values for that marker in all the samples in b1 only. Let us call this b1-mean.
3. Get the difference of the two means, (ref-mean - b1-mean); let us call this difference as mean-scale.
4. Add this mean-scale to every sample of b1 for that marker to make b1-mean equal to the ref-mean.

Table 26.3 explains the rest of the procedure for each marker (or probeset) and for every batch.

Points to Note:

- Un-averaged condition information is used, even if the user has checked the option to 'Average over replicates in condition' during **Create Interpretation**.
- Batch effect correction runs on summarized data set only.

Performing batch effect correction in GeneSpring GX

Start the batch effect correction by using the workflow handle **Quality Control →Batch Effect Correction**. A two-step wizard opens up.

Step 1 of 2 : Batch Effect Correction

Apart from selecting the interpretation, the user can feed the other parameters for batch correction in this step. See Figure 26.9

The parameters required for batch effect correction are

- *Minimum Samples per batch* Default value is 5; user can change this number, based on the data in the experiment.

Table 26.3: Batch Effect Correction

T test	Distribution of samples in a batch against the distribution of the rest of the samples in other batches. Default parameters of 'not against zero, equal variance, un-paired, do not remove outliers' for T test.
P Value	For a batch, the number of p values obtained are equal to the number of markers.
Corrected P Value	Apply Benjamini Hochberg correction to correct for False Discovery Rate (FDR). This yields the corrected p value or q value.
Choosing for correction	For a marker, if the corrected p value is ≤ 0.001 or the specified cut-off, then the batch in which it is present is a candidate for batch correction.
No Correction done if	<ol style="list-style-type: none"> 1. If the batch has less than the defined <i>Minimum Samples per batch</i> 2. If the total number of batches requiring correction is less than '<i>Percentage of bad batches allowed</i>', then none of the batches are subject to correction.
Bad and Good batches	For every marker, get the number of batches where it needs correction ('bad' batches') and the number where it does not need correction (after discounting as per the above rules).
Reference Batch	<ul style="list-style-type: none"> • If, for a marker, all the batches require correction, then the batch with the least number of bad markers is used as the reference batch for correction. In other words, the batch with minimum number of markers with p values ≤ 0.001 or the specified cut-off, is chosen as the reference batch. • In case of only two batches or if the p-value cutoff is chosen as 1, the batch coming alphabetically first will always be chosen as the reference batch, and all the other batches will be corrected taking this batch as the reference. • In case a particular probe needs correction in all the batches, then the first batch among the input batches (sorted alphabetically) with minimum number of corrections will be chosen as the reference batch.

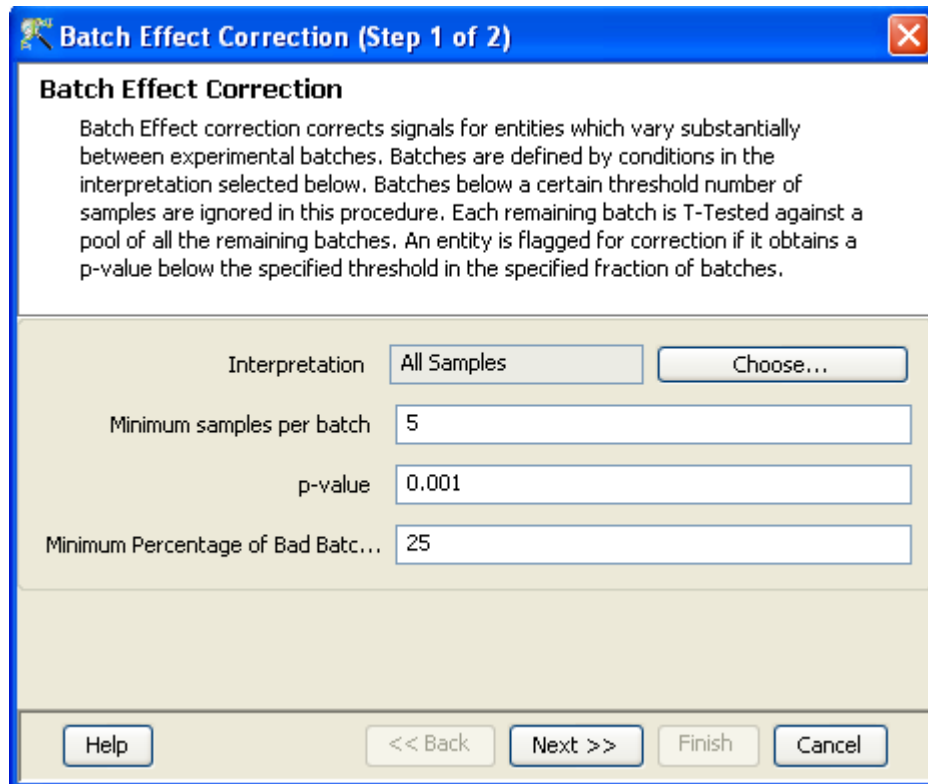


Figure 26.9: Batch Effect Correction - Step 1

- *P value* Default value is 0.001; user can change based on requirement.
- *Percentage of bad Batches Allowed* Default value is 25 percent.

On clicking *Next*, the batch effect correction procedure starts; it proceeds probeset by probeset. If at the end there are no batches that qualify for correction, based on the input parameters, a message is shown that "None of the features was corrected. Try increasing p-value". At this stage, the user can change the parameters in step 1 and try again, or cancel the batch effect correction step.

Step 2 of 2 : Output Views

The result of the batch effect correction is displayed in Spreadsheet, Summary Statistics and Histogram views and is shown in [Figure 26.10](#)

- **Spreadsheet:** Displays the percentage of correction performed to the $\log(A+B)/2$ signals for each entity for each interpretation condition. Negative values signify a decrease and positive values signify an increase. Note that a value of '0.0' indicates 'NO BATCH EFFECT CORRECTION'.
- **Summary Statistics:** For each of the batch, summary statistics are shown here.

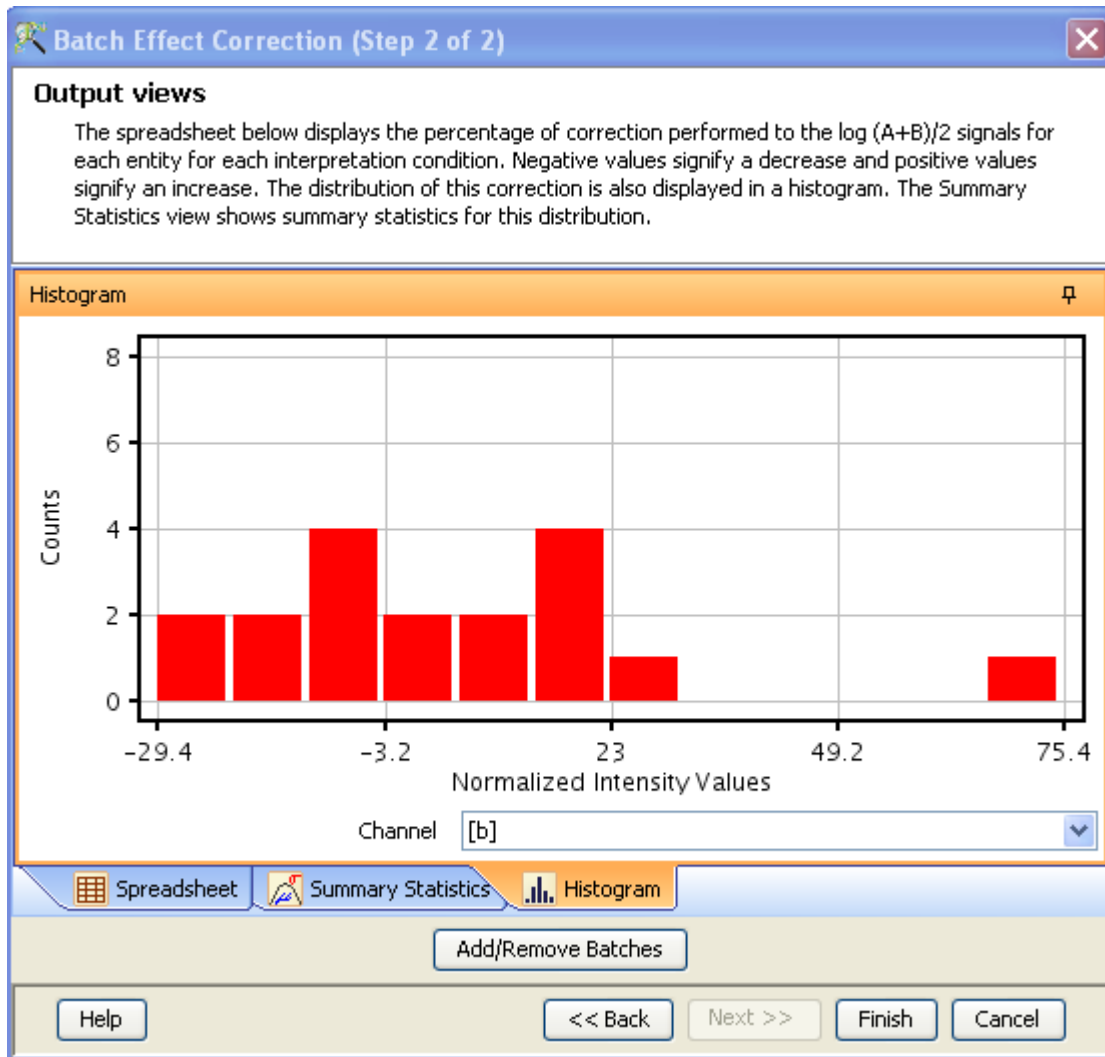


Figure 26.10: Batch Effect Correction - Step 2

- **Histogram:** Distribution of correction is displayed in histogram.

Clicking on the *Add/Remove Batches* brings up a window where batches can be pushed between the two options - 'Batches not to be considered' and 'Batches to Keep'. This is an explicit handle to remove some batches from batch correction procedure, after viewing the results.

Click *Finish* to complete the batch effect correction process. Note that batch effect correction will overwrite any previous results after showing a warning message.

26.3.5 Copy Number Analysis

Copy Number analysis includes computation of Log Ratios, Copy Number, Allele Specific Copy Number, Parent Specific Copy Number and LOH scores and is applicable only for Affymetrix technologies. See section [Technologies supported by GeneSpring GX](#) for details on how the computations are done for various technologies. Copy Number analysis can go through the 'Against Reference' method or the 'Paired Normal' method and this information is deduced from the interpretation.

Inferring method of analysis from Interpretation in Copy Number Analysis

If the user has normal and tumour samples obtained from the same individual, then he will want to run a 'Paired Normal' method of analysis where the tumour samples are compared against the normal samples. If not, the user can decide to run the 'Against Reference' method of analysis where the samples are compared against the standard or custom reference.

To run 'Paired Normal Analysis':

- Create experiment grouping with the parameters 'Condition' and 'Group'
- Under 'Condition' assign samples to be considered as normal as 'Normal'; assign the rest of the samples with any term other than 'Normal'.
- Under 'Group' give identifiers for each group ensuring each group contains at least and at the most, one 'Normal' sample.
- Create an Interpretation with both 'Condition' and 'Group' as experiment parameters and run 'Copy Number Analysis' from this interpretation.

'Against Reference Analysis' will be run if:

- Chosen Interpretation does not contain 'Condition' and 'Group' as parameters.
- Even if the chosen Interpretation has 'Condition' and 'Group' as parameters, if there are no samples assigned as 'Normal' (case sensitive).

<p>NOTE: Note that you can create as many interpretations as you want with varying parameters and choose to run Copy Number analysis on them. Make sure that the analysis is run from that particular interpretation that you are interested in. Also remember that the Copy Number results will be overwritten if the analysis is run again.</p>
--

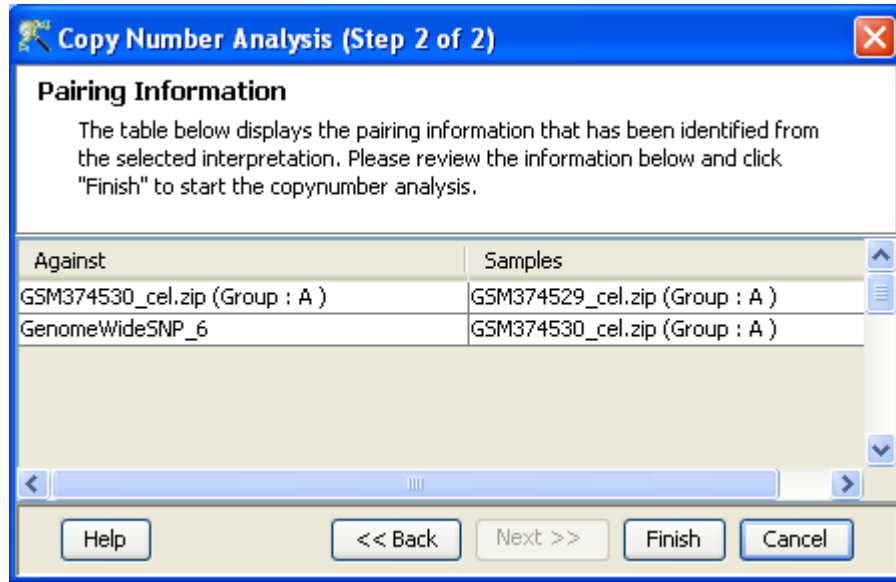


Figure 26.11: Copy Number Analysis - Paired Normal Method

The computations take different course depending on whether it is 'Against Reference' or 'Paired Normal' type and also depending on the array type and is explained in the section [Technologies supported in GeneSpring GX](#).

Copy Number Analysis can be started from the workflow menu **Copy Number Analysis**.

A 2 step wizard comes up.

Step 1: Interpretation Chooser Choose the interpretation that you want the Copy Number analysis to be run on. See section [Significance of Interpretation in Copy Number analysis](#) for detailson how to define interpretation for running 'Against Reference' analysis and 'Paired Normal' analysis.

Step 2: Pairing Information Based on the interpretation, **GeneSpring GX** infers if the analysis is against reference or paired normal and shows the pairing information here. User can go back and choose a different interpretation if this is what he wants.

The Figure 26.11 shows an instance where 'Paired Normal' method of analysis is inferred. Note that in this type of analysis, the tumour samples are compared against the normal samples while the normal samples are compared against the standard / custom reference.

The Figure 26.12 shows an instance where 'Against Reference' method of analysis is inferred. Whatever reference was chosen during experiment creation (standard or custom) will be used here.

NOTE: Only one Copy Number / LOH value will be saved per experiment. If the user runs the analysis again, the previous values will be over-written.

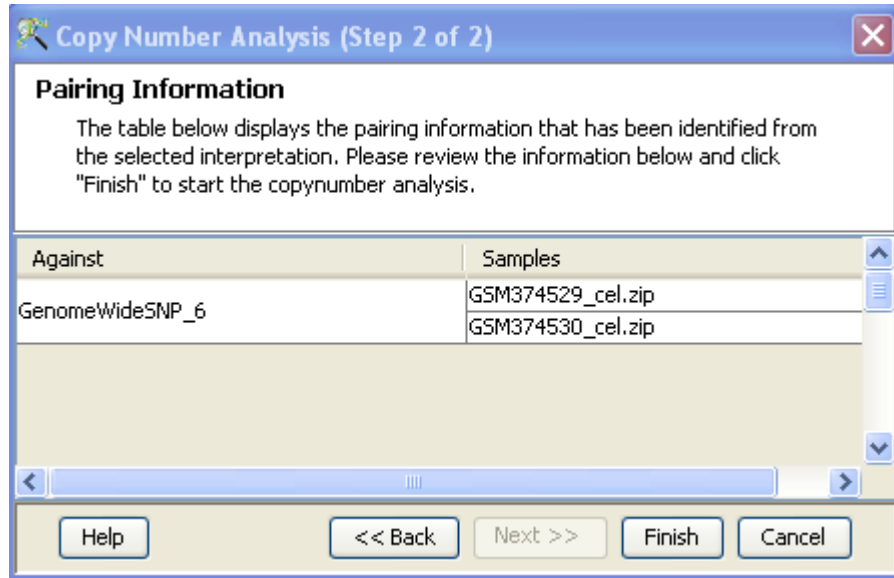


Figure 26.12: Copy Number Analysis - Against Reference Method

The Copy Number analysis results can be viewed as a [Heat map](#) or in the [Genome Browser](#). Further analysis like [Common Genomic Variant Regions](#) or [Filters](#) or [Results Analysis](#) can be carried out.

26.3.6 Common Genomic Variant Regions

This involves identification of regions of the genome that are significantly amplified or deleted across a set of samples. The ultimate motive is to relate these regions of aberration to cancer pathogenesis and the method is commonly referred to as **Genomic Identification of Significant Targets in Cancer (GISTIC)**.

GISTIC analysis runs on the outputs of Copy Number Analysis; it runs independently on the regions of Copy Number Amplification and Copy Number Deletion. At the end, two genomic regions which show aberrations with respect to Copy Number amplification and deletion are identified. Table 26.4 gives a snapshot of GISTIC process.

Table 26.4: snapshot of 'Common Genomic Variant Region' Detection Algorithm

Inputs	Genome of each sample with Copy Number data and Log Ratios; Illumina outputs can be directly used for GISTIC analysis. For Affymetrix CEL files, run Copy Number analysis and then run GISTIC.
Modes of running GISTIC	GISTIC can run in Coarse or Fine mode. Coarse mode applies Q value cut-off filter to identify regions of aberration. In Fine mode, after identifying peaks, aberrant segments are assigned to each peak; then all aberrant segments corresponding to this peak are set to '0' and iteratively, we detect further peaks till we reach the threshold for the peaks or till we fail to detect any more peaks. (Q value obtained by this process is called Peeled Q value.)
Output	<ol style="list-style-type: none"> 1. A gene list called 'Overlapping genes with identified regions of aberration at the specified Q value cut-off'. This gene list will be absent if no genes are found in those regions. In Fine Mode, the region that is passed for identifying overlapping genes is defined by the peak position. In Coarse mode, the entire region identified as region of aberration after applying Q value cut-off filter is passed for identifying overlapping genes. 2. Probe set wise Q value list (unfiltered) 3. Probe sets in identified regions of aberration (after applying the Q value cut-off filter); the Q value listed here is the peeled Q value.
Configurable Options	Many GISTIC parameters can be set from the menu Tools → Options → Copy Number Algorithms → GISTIC . See section Configuration options for Copy Number analysis .
Reference	Broad Institute's Cancer Program Publication [23]

Process:

Obtaining Mean Log Ratios Segment the genome based on Copy Number; for each segment with same Copy Number, get the mean of the log ratios and replace the log ratio value of each probe with this mean log ratio value.

Identifying Regions of Amplification Look at the Mean Log Ratio value for each segment for each sample; by default, any value above 0.9 is retained as it is and all those below 0.1 are thresholded at 0. The lower and upper cut-off values are configurable.

Identifying Regions of Deletion Look at the Mean Log Ratio value for each segment for each sample; any value below -1.3 is retained as it is and all those above -0.1 are thresholded at 0. The lower and upper cut-off values are configurable.

Obtaining G score Get the average of sum of the log ratios for each segment (done individually for regions of amplification, deletion and LOH) to get G score for each probeset across all samples.

Obtaining P value From the log ratios for all probe sets for all samples, find frequency by dividing by binning interval. Multiple this frequency across samples and this will give a distribution curve. Each G score is mapped to an index on the frequency curve. The area of the curve beyond this index divided by the total area under the curve gives the P value for each score.

Obtaining Q value and Peeled Q value Apply Benjamin Hocheberg Correction to obtain Q value which is corrected P value. Peeled Q value is the minimum Q value in the Region considering only the aberrant segments associated with this Peak.

Assigning Peak Q value Peak Q value is the minimum Q value in the Region considering only the aberrant segments associated with the Peak. This is shown for both Coarse and Fine mode ; Fine mode shows Peeled Q value. When there are multiple segments associated with a probeset, one Peak Q value corresponding to the most significant peak is shown.

Handling X and Y Chromosomes Due to Copy Number corrections, it is possible that large number of deletions are detected in X and Y chromosomes; hence these are not considered by default for GISTIC analysis in both Coarse and Fine mode.

Running Common Genomic Variant Region analysis

Common Genomic Variant Regions analysis is enabled only for Copy Number experiments and can be called from the workflow menu **Analysis** → **Common Genomic Variant Regions**. Note that the inputs required for the analysis are Log Ratios and Copy Number segmented Genomes. The analysis is supported for both Affymetrix technology and Illumina technology. In the case of Affymetrix technology, it can only run after running Copy Number analysis as the required inputs are generated only after Copy Number analysis.

NOTE: For Affymetrix technology, Common Genomic Variant Regions needs Copy Number analysis results to proceed. For Illumina technology, after creating the experiment, it can be directly called.

A 4-step wizard opens up on clicking .

Step 1: Interpretation Chooser Choose an interpretation; samples for which the analysis will run will be determined by the interpretation and it will be shown in this step.

Step 2: Q Values This step gives the unfiltered probe sets along with Q values for all the regions of Copy Number amplification and deletion. There are tabs at the bottom of the table for viewing 'Amplification Q values' and 'Deletion Q Values'. These lists are saved and shown as 'Unfiltered Q value' in a folder called 'Gistic results' in the experiment.

In this step, there is an option to choose 'Coarse mode' or 'Fine mode' for identifying regions of aberration. Coarse mode applies just the Q value cut-off filter; In Fine mode, after identifying peaks,

Step 3: Identified Regions

Identified Regions
This page shows significant regions which pass the peak q-value cut-off specified below. Regions for Amplification and Deletion are shown separately

Deletion Regions 73 Rows

Region id	Chromos...	Segment Start	Segment ...	Focal / Br...	QValues	Peeli
Region1	chr1	102472018	102622376		0.132338	0.1
Region2	chr1	106576334	107185261		0.161772	0.1
Region3	chr1	145235531	145347239		0.0581641	0.05
Region4	chr1	145390602	145484985		0.101717	0.1
Region5	chr1	167424543	167456395		0.108687	0.1
Region6	chr1	170755160	171265834		0.247815	0.2
Region7	chr10	13809953	14020191		0.225421	0.2
Region8	chr10	85144905	85241723		0.120355	0.1
Region9	chr10	113332410	113381547		0.0757998	0.07
Region10	chr11	18907033	18918255		0.0325332	0.03
Region11	chr11	29369017	29607875		0.179348	0.1
Region12	chr11	88664103	88781154		0.0399775	0.03

Amplification Regions 1 Rows | Deletion Regions 73 Rows

Default Q-Value cutoff: 0.250 [Recompute Regions]

[Help] [<< Back] [Next >>] [Finish] [Cancel]

Figure 26.13: Common Genomic Variant Regions - Step 3

aberrant segments are assigned to each peak; then all aberrant segments corresponding to this peak are set to '0' and iteratively, we detect further peaks till we reach the threshold for the peaks or till we fail to detect any more peaks. This is called Peeled Q value.

Figure 26.13 shows the identified regions with a default cut-off value of 0.25 for Amplification, Deletion and LOH scores. The result is in the form of a table giving information on Chromosome number, Segment Start and End, whether it is Focal or Broad region along with the Q value and Peak position for that region. These are basically regions of aberration listed for amplified/deleted Copy Number segments.

In Fine Mode, the peak region is identified considering only the aberrant segments associated with the Peak. Peeled Q value is the minimum Q value in the region and this peeled Q value is reported as the peak Q value for fine mode.

It is possible to iteratively redo the computation with a different Q value cut-off value from this step.

Step 4: Overlapping Regions The probes corresponding to the identified regions of aberration are identified and the genes containing these probes are then output as 'Overlapping genes' provided such genes exist. This gene list is created for each of amplification and deletion provided such genes exist, along with probe-wise Q value list. See Figure 26.14

NOTE - Rerunning with same parameters overwrites results.

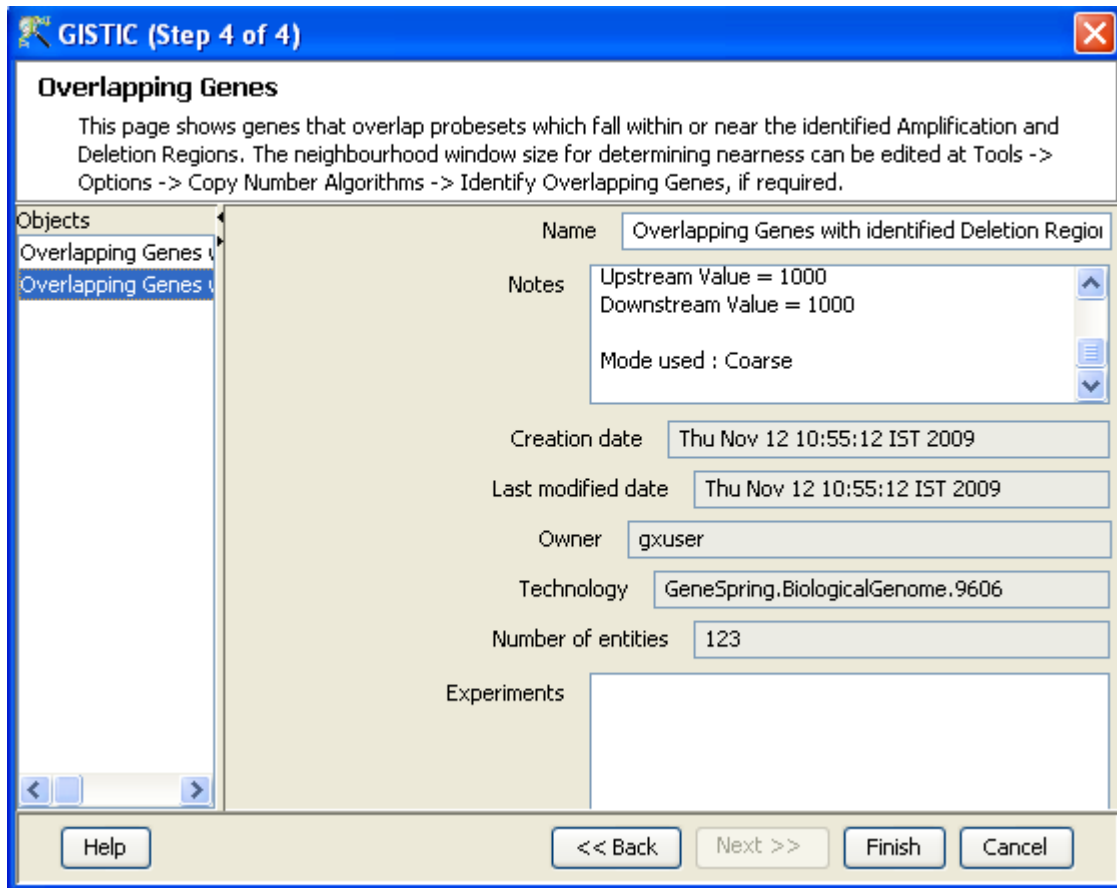


Figure 26.14: Common Genomic Variant Regions - Step 4

Note on Identifying Overlapping Regions in GISTIC

Given an entity list, **GeneSpring GX** identifies overlapping genes using a default value of 1000 units to detect genes upstream and downstream, as explained in section [Identify Overlapping Genes](#). When we run 'Identify Overlapping Genes' from within GISTIC, the entity list that is input is slightly different for Fine Mode, compared to the Coarse Mode. In Fine mode, the entity list corresponding to only the peak region is used for identifying overlapping genes. Whereas in Coarse mode, the entity list corresponding to the entire region identified as aberration is used for identifying overlapping genes.

26.3.7 Filters

Once the Copy Number analysis is done, various filters can be applied to pull out data satisfying certain specified criteria; typically the results of filtering are entity lists. Filtering is possible on Log ratio, Copy-number and Copy Number confidence, LOH score, Parent Specific Copy Number and Known Genomic Variant Regions (based on DGV).

Workflow options include:

1. [Filter By Regions](#)
2. [Identify Copy Neutral LOH](#)
3. [Filter by Parent Specific Copy Number](#)
4. [Filter by known CNVs](#)

Filter By Regions

'Filter by regions' is a top level filter where ranges of data can be specified for Copy Number (CN), Copy Number Confidence, LOH and Log ratios.

- Inputs:**
1. Entity list - 'All Entities' by default
 2. Interpretation to derive the condition information
 3. Sample aggregation information
 - x out of y samples should satisfy the filter criteria
 - x % of samples in any y out of z conditions should satisfy the filter criteria
 4. Minimum number of probesets in the region
 5. Filter criteria on the various fields with 'And' condition.

- Process:**
1. Go through each probeset in sequence along the genome. At each probeset, determine if the probeset satisfies the specified Copy Number (includes Copy Number, Copy Number confidence, log ratios) and LOH conditions. If the probeset does not have values for a particular condition (e.g., LOH for CN probesets) then inherit the value for that flag from the previously measured value. This is applicable only for LOH - any other missing value is ignored.
 2. For each sample, consider only those probesets which satisfy both the CN and LOH filter conditions.
 3. Aggregate the probe sets across samples and apply the sample aggregation condition on probesets resulting from step 2. For each such probeset, we now get a pass/fail indication depending upon whether or not the sample aggregation condition is satisfied.
 4. On the aggregated list, apply the region size threshold and output only those probesets, which obey this filter.

- Outputs:**
1. **Output view:** A searchable spreadsheet with the entities that passed the filter criteria will be displayed
 2. **Entity list:** An entity list containing all SNPs which pass the filter criteria is saved.

Performing 'Filter by Regions' in GeneSpring GX : From the Copy Number experiment, launch 'Filter by Regions' from the workflow browser →Filter →Filter by Regions. A 3 step wizard opens up.

Step 1: Input Parameters Takes the 'All Entities' list by default and the selected interpretation for filtering.

Step 2: Filter Conditions Takes the sample aggregation information criteria and the filter conditions. For sample aggregation, choose to retain entities where

- A defined number of samples out of the total samples meet the filter conditions.
- Or a defined percentage of samples in the specified number of interpretations meet the filter conditions.

Option to define minimum probesets per regions meeting the filter conditions is also there.

Filtering itself can be done on Copy Number values, Copy Number Confidence, Log ratios and LOH scores. Check the required fields and define the condition; Conditions include 'equal to, Not equal to, Less than, Greater than, Less than or equal to, Greater than or equal to, in the range and outside the range'. Define the relevant Filter value after choosing the condition. Remember to 'Enter' after giving the filter value.

See Figure [26.15](#)

Step 3: Output Views of Filter by Region Shows the probe set IDs and the number of entities that passed the filtering criteria.

Step 4: Save Entity List Displays the filtered entity list along with details like probe set ID, Affy SNP ID, dbSNP RS ID, Chromosome number, Physical position and Strand information. *Finish* will exit the wizard and create an entity list called 'Filter on Regions' under 'All Entities' in the experiment.

Identify Copy Neutral LOH

This filter is applied only on the copy neutral regions of the genome within the specified LOH threshold.

Process

1. For each sample, go over all probesets to determine regions of contiguous probesets which have LOH greater than the specified threshold. This threshold can be set from the menu **Tools** → **Options** → **Copy Number Algorithms** → **Segment Thresholds** and the default value is 0.5 (Value above 0.5 indicates Loss of Heterozygosity).
2. If the probeset does not have LOH value (e.g., LOH for CN probesets), then inherit the value from the previously measured value.
3. For each of the regions identified using the above process, find out the average copy number. The average copy number of the region is the mean of the copy number values of all the probesets in the region.
4. If this average copy number is in the range of 1.5 to 2.5, then retain all the probesets. Else the entire region is said to have failed the filter.

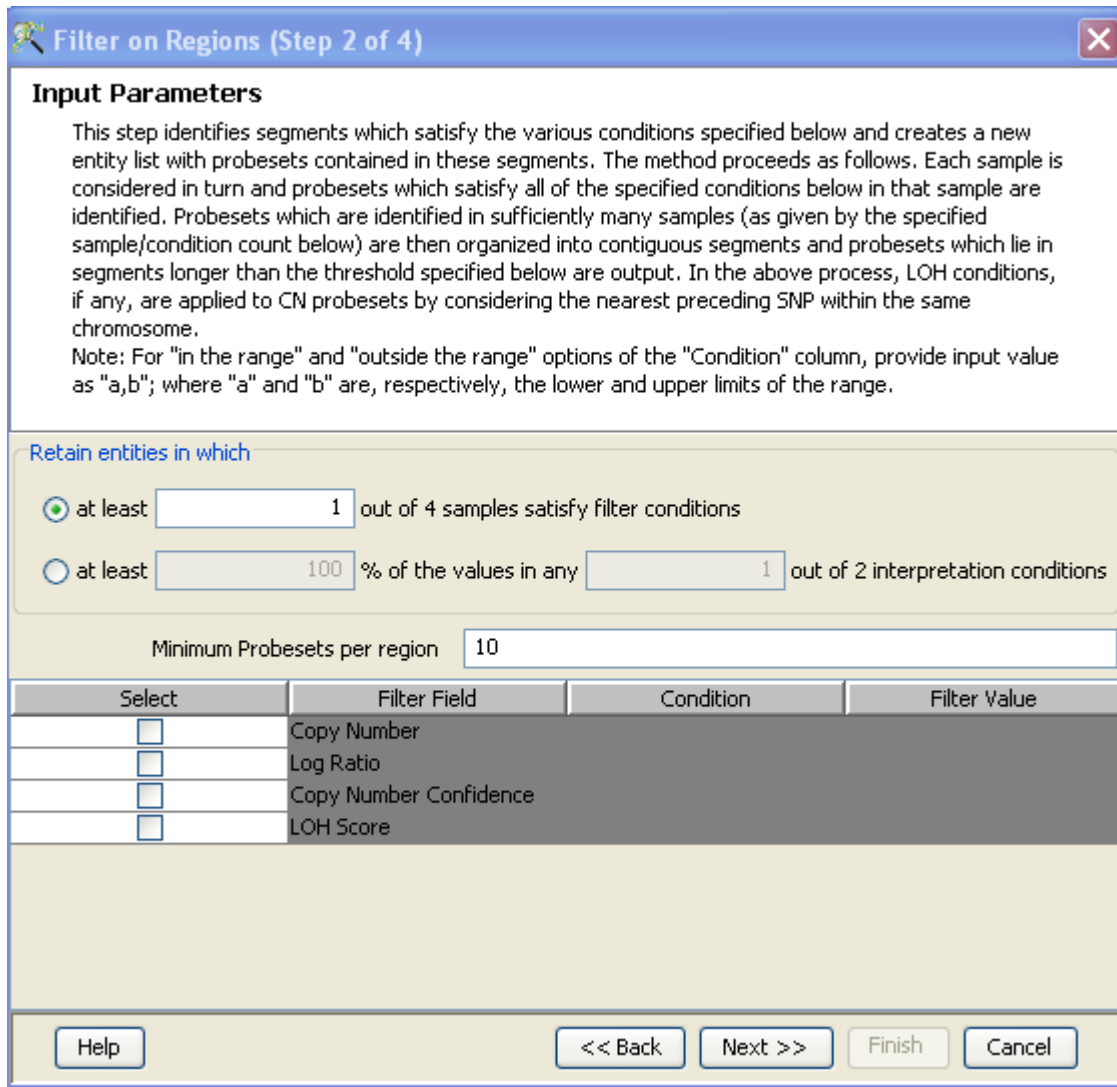


Figure 26.15: Step 2: Filter Conditions for Filter by Region

5. Aggregate the probe sets across samples and apply the sample aggregation condition on probesets. For each such probeset, we now get a pass/fail indication depending upon whether or not the sample aggregation condition is satisfied.
6. On the aggregated list, apply the region size threshold and output only those probesets, which obey this filter.

In human genome, the normal level of homozygosity is around 70 % and are confined to less than 100 kb. When doing an 'Against Reference' analysis, any region with more than 70 % homozygosity can be safely assumed to be showing LOH. This combined with some meaningful filter conditions (using the option 'Minimum probesets per region') can reduce false positives and remove small pockets of high homozygosity while picking up LOH regions.

In V6 experiments, SNP-SNP distance is typically 1.3kb and this translates to a minimum of 75 probesets if one wants to apply the 100kb filter.

Parent Specific Copy Number Filter

Filter criteria is applied on parent specific Copy Number; the user can define the desired Copy Number value/range for allele A and allele B in the filtering criteria in addition to the sample aggregation information. Both SNP and CN probes are considered for PSCN filter.

A probe-set may show missing PSCN value due to the following reasons:

1. It is a CN probeset.
2. SNPs may show missing PSCN because they were part of a long homozygous stretch and hence ignored while assigning PSCN values.
3. SNPs may also show missing PSCN because its copy number and subsequently ASCN was not calculated for reasons like missing birdseed clusters.

These missing PSCN values are inherited from previously available values, as explained in section [Filter by Regions](#).

Note that this filter is not supported for Affymetrix 50/100k array set.

Go to workflow **Filter** → **Filter by Parent Specific Copy Number**; a 4 step wizard will open up.

Step 1: Interpretation Chooser Select an interpretation for the filtering.

Step 2: Input Parameters Sample aggregation information can be defined in terms of:

- Number of samples satisfying the filter criteria (OR)
- Percentage of samples in a defined number of interpretations satisfying the filter criteria

Minimum probesets per region can also be defined.

See Figure [26.16](#)

Step 3: Output Views on Parent Specific Copy Number calls Gives the Probe set IDs and the number of samples for which the entity passed the filter criteria.

Step 4: Save entity list Details of the entity list created after applying filter conditions are shown here. Clicking *Finish* will create a new entity list called 'Parent Specific Copy Number Filter' in the experiment navigator and exit the wizard.

Filter on Parent Specific Copy Number Difference (Step 2 of 4)

Input Parameters

Parent Specific Copy Number denotes the number of chromosomal copies for each of the two chromosomes. All entities which are part of PSCN segments with the following property will be retained: the segment should have at least the specified number of entities and should match the PSCN difference condition specified below in the specified number of samples. PSCN difference here denotes the difference in the number of copies between the two chromosomes.
 Note that homozygous stretches in the chromosome will have missing PSCN values and will not be considered in the computation below.

Retain entities in which

at least out of 4 samples satisfy filter conditions

at least % of the values in any out of 2 interpretation conditions s:

Minimum Probesets per region

Filter Field	Condition	Filter Value
PSCN Difference	=	

Help << Back Next >> Finish Cancel

Figure 26.16: Step 2: Input parameters for PSCN Filter

Filter By Known CNVs

Process is explained in Table 26.5

On clicking the workflow menu **Filter** → **Filter by Known CGVs**, a 2 step wizard opens up.

Step1: Entity List Chooser Pass an entity list here

Step 2: Results Shows the number of entities in known genomic variant regions as well as the number of entities outside this region. Clicking *Finish* will create two entity lists in the experiment navigator - 'Entities in known genomic variant regions' and 'Entities outside known genomic variant region'.

CGVs	Common Genomic Variant Regions.
Inputs	Entity list and the Database of Genomic Variants (DGV) packaged within GeneSpring GX . See http://projects.tcag.ca/variation/ for information on DGV.
Process	For each probeset in the input entity list, identify if it overlaps any of the known genomic variant region.
Outputs	Entity list with all entities that overlap any of the known genomic variant regions

Table 26.5: Filter by CGVs

26.3.8 Views

For Copy Number analysis, views are supported in the form of heat map and the Genome Browser.

Heatmap

For the given entity list and interpretation, heat map can be drawn for Copy Number and LOH scores for each sample. See section [Heat Map](#) for complete details on heat map.

See Figure [26.17](#) for a Heap Map view for a Copy Number Experiment.

An additional feature on the heat map for Copy Number results is the functionality to create entity lists from selection. Click the icon **Create Entity list** in the toolbar; it creates an entity list with the selection on the heat map. The entity list called 'Region map entity list' is added to the experiment navigator.

For Copy Number analysis, the heat map is actually a region map as it shows the Copy Number or LOH regions. On **Right click** → **Properties** from the heat map, there is an option to choose 'Probeset ID, Chromosome or Physical Position' under 'Visualization' tab. Each probe or chromosome or physical position is samples and the view can be collapsed / expanded to fit all. 'Column' tab allows choosing samples to view. There is a chromosome browser at the bottom.

Genome Browser

Genome Browser allows viewing Copy Number data at the genome level, and in combination with gene/transcript annotation data of the organism. The data can be viewed along with data from other technologies (expression, etc) for the same organism. See chapter [Viewing Copy Number Experiments in Genome Browser](#) for details.

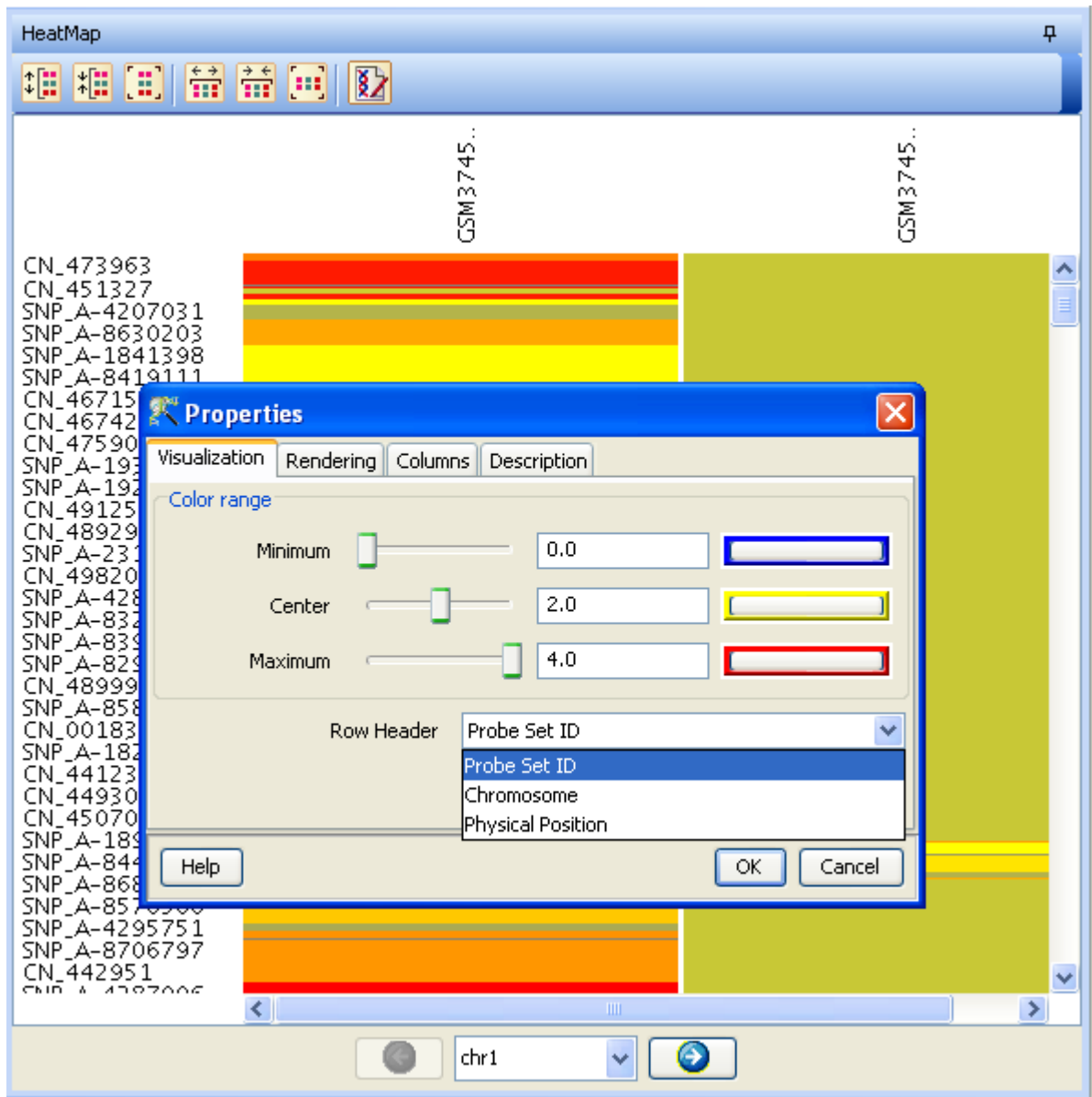


Figure 26.17: Heat Map View for a Copy Number Experiment

26.3.9 Results Analysis

Identify Overlapping Genes

Helps to identify the genes present in regions of interest (detected from Copy Number analysis); Process explained in Table [26.6](#).

Start 'Identify Overlapping Genes' from the workflow **Results Analysis** → **Identify Overlapping Genes**. A 2 step wizard opens up.

Step 1: Entity List Chooser Pass the entity list here.

Step 2: Output View Gives a list of the overlapping gene IDs and the chromosomes they are located on; the upstream and downstream region to search for genes is a configurable option from the menu **Tools** → **Options** → **Copy Number Algorithms**. Clicking *Finish* will create the entity list in the experiment navigator called 'Overlapping Genes' and exit the wizard.

For description of 'Identifying Overlapping Genes' with GISTIC result, see section [GISTIC - Identify Overlapping Genes](#).

GO Analysis

Once a gene list is available from Copy Number analysis (generated by running 'Identify Overlapping Genes' OR 'GISTIC'), one can do GO analysis. For details, see section [Gene Ontology Analysis](#).

Pathway Analysis

Pathway analysis is also possible with a gene list. See section [Pathway Analysis](#) for details.

26.3.10 Utilities

The utilities section in Copy Number analysis allows user to perform some tasks relevant to downstream operations of Copy Number analysis quickly.

Table 26.6: Identify Overlapping Genes

<p>Inputs</p>	<ol style="list-style-type: none"> 1. An entity list. 2. Upstream and downstream distance to look for genes in the chromosome. Default value is 1000 units on both sides (start and end locations); this value can be changed from the menu Tools → Options → Copy Number Algorithms → Identify Overlapped genes - Upstream / Downstream Values.
<p>Process</p>	<ol style="list-style-type: none"> 1. Identify the valid probesets from the selected entity list. A valid probeset is defined as one which has both the chromosome and physical position information. 2. For all probesets belonging to a chromosome, chr1, get the position map from the interval group map. Using the chromosome, start and end position columns in the biological genome, an interval group map is generated on the fly. 3. For each probeset use the physical position information to get a list of indices of overlapping genes, by applying the upstream and downstream distances. 4. Use these indices to get the gene-ids from the biological genome. 5. These gene-ids are the overlapping genes.
<p>Outputs</p>	<ol style="list-style-type: none"> 1. For the 'Identify Overlapping Genes' step, a spreadsheet with one column as the gene-id and another column with the chromosome number is shown to the user. 2. An entity list where each entity is a gene is saved at the end.

Import Entity List from File	A simple import functionality by matching column names. See section Import Entity List from File for details.
Create Probe List	This utility allows searching an existing entity list for certain positions of chromosomes. The result is a SNP list which match the search criteria, based on physical coordinates.
Filter on Entity List	This utility allows user to filter an Entity list using its annotations and list associated values. See section Filter on Entity List for details.
Export Segmentation Results	User can choose to export Copy Number or LOH segments as also segments identified to have aberrations (amplifications and deletions). GeneSpring GX 11.0 automatically gives the name of the cytoband the segments overlap with. To get the cytoband overlap information, user should have downloaded the annotation data using the menu Annotation → Update Genome Browser Data → From Agilent Server . The tool prompts for choosing an organism build to obtain this information. The exported text file can then be dragged and dropped into Genome Browser for viewing. See chapter Viewing Copy Number Experiments in Genome Browser for details.

Table 26.7: Utilities in Copy Number Analysis

26.4 Copy Number analysis of Illumina

Users can directly bring in Copy Number results from Illumina’s GenomeStudio within **GeneSpring GX** and do analysis. There is a plug-in packaged with **GeneSpring GX** 11.0 to extract output from Illumina for analysis in **GeneSpring GX** .

26.4.1 Obtaining Data from Illumina

The following steps will guide in importing genotype information from GenomeStudio:

1. Copy the plug-in from `INSTALLDIR\app\Illumina\GX.Genotyping.Export.dll` to `Genomestudio\modules \ BSGT \ ReportPlugins\`.
2. Restart GenomeStudio.
3. Open the Project.
4. Select **Reports** option from the the **Analysis** menu.
5. Click **Report Wizard** option from the **Report** sub menu. This launches the **Report Type** wizard page.

6. Select **Custom Report** radio button, and **GeneSpring Exporter 1.0** from **Strand Life Sciences** from the drop-down list, and click *Next*. This launches the second page of the **Sample Groups** wizard page.
7. Select the sample groups you want to include in the report, and click *Next*. This launches the **Destination** wizard page.
8. Provide an **Output Path** (you can use the **Browse** button), and **Report Name**, and click **Finish**. This launches a **Progress bar** with information about the incremental progress in generating the Report. After the **Report** is written completely, you will see a message with an option to view the **Report**.

The plug-in will extract the following information relevant to the technology:

- dbSNPID
- Chromosome Name
- Chromosome Position
- HWE (for association experiments)
- CNV regions

From the Illumina output files, **GeneSpring GX** directly uses the following values for each sample.

- log R - Provides the log (base 2) ratio of the normalized R value for the SNP divided by the expected normalized R value.
- BAF (B allele Frequency)
- Copy Number values and their confidences
- Genotype of the subject SNP for the sample along with the score.

NOTE: Ensure that Copy Number analysis is done in GenomeStudio before running the plug-in. All the columns required for analysing Illumina data will be available only after running the Copy Number analysis within GenomeStudio.

Allele specific Copy Numbers are computed for Illumina outputs within **GeneSpring GX** from BAF and Copy Number values. Parent Specific Copy Numbers are computed as explained in section [PSCN](#).

$$BAF = \frac{B}{A + B}$$

B is the B allele. In place of A+B, the Copy Number value is substituted and the BAF value is used on the left hand side of the equation to obtain the value of 'B'. A is obtained from the total Copy Number and the B value. Thus the allele specific Copy Number values are obtained from Illumina output.

Note that log R is used directly for Illumina wherever log Ratio is mentioned.

26.4.2 Handling Missing Values

- If any missing values are present either in Chr or Position columns, those entities won't be considered in copynumber datasets/analysis.
- If any missing values are present in any sample columns, those values are treated as missing values.

26.4.3 Workflow description for Illumina Outputs

To start an Illumina experiment, user needs to have run the plug-in and extracted output from GenomeStudio as explained in the section above on '**Obtaining Data from Illumina**'.

Workflow remains almost the same for Illumina outputs as with Affymetrix files explained above. Some significant differences are:

1. Technology is created on the fly based on the samples.
2. QC step is skipped for Illumina.
3. Since Copy Number information is already present in Illumina, the analysis part skips this step and instead calculates only 'Allele specific Copy Number'.
4. There is no filter for 'Copy Neutral LOH' as LOH values are not present.
5. It is not possible to filter on entity list.

The table below summarizes the steps in Illumina workflow highlighting how it is different from the Affymetrix workflow explained in Table 26.8.

26.5 Create Custom Reference

GeneSpring GX allows advanced user to choose their own references for comparison instead of using the standard 270 HapMap samples, as explained in section [Technologies supported in GeneSpring GX](#).

Create Technology	Technol-	Technology is created on the fly based on the samples used to create the experiment.
Create Experiment	Experi-	Experiment creation is similar to Affymetrix; use the output files generated from GenomeStudio Refer section Create Experiment for other details.
Experiment Set up	Set	As explained for Affymetrix files in section Experiment Set up .
QC		QC step is skipped for Illumina as GeneSpring GX works with outputs from Illumina's Genome Studio only.
Analysis		Computes Allele specific Copy Number as explained in section Allele Specific Copy Number for Illumina Outputs . Common Genomic Variant Region analysis (GISTIC) also runs as explained for Affymetrix files in section Common Genomic Variant Regions .
Filters		Other than the filter for identifying copy neutral LOH regions and general filtering on LOH, all filters that run on Affymetrix files work on Illumina outputs; refer section Filter .
Views		Heap Map and Genome Browser
Results analysis		As explained for Affymetrix files in section Results Analysis .
Utilities		Remain the same as in Affymetrix files Utilities except for the utility to filter on entity list. This is not supported for Illumina

Table 26.8: Workflow for Illumina output files

Users can create their own custom references for running Copy Number analysis within **GeneSpring GX** from the menu **Tools** → **Create Custom Reference**. Ensure that you have downloaded the required technology before proceeding for custom reference; See [Create Technology](#) for details. Note that custom reference is relevant only for Affymetrix technologies and not for Illumina Copy Number outputs.

See section [Technologies supported by GeneSpring GX](#) for details on how the reference gets created for various Affymetrix technologies.

On clicking **Create Custom Reference**, a 3 step wizard opens up.

Step 1: Load Data Choose files/samples to be used for custom creation; note that you cannot mix files from different technologies.

Step 2: Pair CEL files Shows interface for pairing CEL files.

Step 3: Reference Description Allows user to define a name and notes for the custom reference.

26.6 Useful information

26.6.1 Using disc cache

Under the menu **Tools** → **Options** → **Copy Number Algorithms** → **APT Execution Options**, 'Use Disc Cache' is an option. It is advisable to use the disc cache in Windows system. Note that it will fail with Linux Operating System with more than 90 samples. When disc cache is not used, the execution process creates batches of probesets and this can slow down the process. In worst situation, the process can quit if there is memory outage. Using disc cache helps to avoid these problems and uses the disc memory while running processes.

26.6.2 Entity Lists and Translation rules in Copy Number

Entity lists in Copy Number analysis behave as normal entity lists generated elsewhere in the tool, for example, in expression experiments. Right click operations on entity lists are valid except for 'Search' option in the entity list inspector; there is no 'search' option on entity lists created in Copy Number experiments.

Translation of entity lists created in Copy Number and association experiments is possible only within these technologies. It is not possible to translate entity lists from expression to Copy Number experiments or vice versa. By default, Translation is done on the basis of gene IDs; in this case, it will be possible to run translation only between experiments of same technology within Copy Number (for example, only between two V6 experiments). There is also an option to do translation on the basis of **dbSNP RSP IDs** which will allow translation between experiments of all types in Copy Number and even from association experiments, if dbSNP RSP ID column is present.

Note that the explicit translation mapping from the menu **Tools** → **Miscellaneous** → **Translation Mapping** does not work with Copy Number and association experiments.

26.6.3 Configuration options for Copy Number analysis

From the menu **Tools** → **Options** → **Copy Number Algorithms**, various configuration options are available for Copy Number analysis.

GISTIC:

1. **Maximum peaks/chromosome:** This defines the number of peaks to be taken per chromosome starting with the minimum Q value. Default value is 250 for amplification and 250 for deletion regions.

2. **Lower Amplification Cut-off:** Set to 0.1; those below 0.1 would be taken as 0.
3. **Upper Amplification Cut-off:** Set to 0.6; those above 0.6 would be taken as 0.6; those in-between 0.1 and 0.6 would be retained as it is.
4. **Lower Deletion Cut-off:** Set to -1.0; those below -1.0 would be taken as -1.0
5. **Upper Deletion Cut-off:** Set to -0.1; those above -0.1 would be taken as 0.
6. **Use only segments with Copy Number confidence:** Set to 2.0 by default, which indicates low confidence.
7. **Q value cut-off for peak detection:** Set to 0.25 by default. It can also be changed while running GISTIC in step 3.
8. **Interval for binning:** Set to 0.001 by default. This can be increased based on the number of samples.
9. **Normalization-Validation: reduce each sample median to zero:** By default, this is checked.

Options for running CBS:

1. **Validate Change Points using :** Gives option to choose T test or Permutation test for validation the segmentation. Permutation test is slower and takes about 30 minutes per sample. T test is faster.
2. **Number of Permutations:** Applicable only if permutation test is chosen.
3. **p-value cut-off:** Applicable for T test where the change points are validated. Default value is 0.002 but can be changed by the user from this option.
4. **Mean Difference between Segments:** User can also set a mean log ratio difference threshold between segments. Segments which differ in their mean log ratio by a value lesser than this cut-off will be joined.
5. **Minimum length of the Segment** User can set the minimum length of the segment that is acceptable as output from CBS.
6. **Offset to jump during optimization:** In each set of log ratios, algorithm searches for the maximum statistic by considering the data as a circle and by iterating within arcs of length 2 to $(n-1)/2$. The offset defines the arc interval for the search; for the default value of 3, the algorithm will go in steps of 2, 5, 8, After it identifies an arc of length l with the maximum statistic, it checks for $(\text{offset} - 1)$ values above and below this length to get the best estimate of the maximum statistic.
7. **Size of segment to use jump offset:** Defines size of segment for the above condition.

Segment Thresholds: Allows defining the LOH threshold to be considered while applying filtering and GISTIC analysis.

Fawkes Options:

1. **Number of parallel threads:** For optimization in multi core machines.
2. **Number of probes in the homozygous stretch:** For Parent Specific Copy Number computation, user can define the number of probes to be present in a segment to qualify as homozygous stretch.
3. **Minimum Number of probesets in a segment:** For Parent Specific Copy Number computation, this sets the number of probes to qualify as a heterozygous segment.

Identifying Overlapping Genes - Up Stream and Downstream values: Set to 1000 units.

APT Execution Options: Allows user to 'Use Disc Cache'. It is advisable to use the disc cache in Windows system. Note that it will fail with Linux Operating System with more than 90 samples. When disc cache is not used, the execution process creates batches of probesets and this can slow down the process. In worst situation, the process can quit if there is memory outage. Using disc cache helps to avoid these problems and uses the disc memory while running processes.

26.6.4 Performance Statistics for Copy Number Analysis

For Copy Number and Association Experiments with more than 100 samples, a 64-bit QuadCore Desktop with more than 4Gb RAM is recommended.

Approximate statistics for 75 samples of Affymetrix Genome-Wide Human SNP Array 6.0 on a 32-bit, QuadCore Desktop with 2 Gb RAM are:

- Experiment Creation takes 1.5 hours
- Combined time for the following operations for 'Against Standard Reference' analysis is 7-8 hours:
 - [Log Ratio](#)
 - [CBS segmentation to identify Copy Number segments](#)
 - [Copy Number computation which is post processing of CBS segments](#)
 - [Parent Specific Copy Number Computation](#)
 - [LOH analysis](#)
- Paired Normal analysis for the same experiment could go upto 14-15 hours depending on the pairings.

26.7 Copy Number Algorithms

26.7.1 BRLMM

GeneSpring GX calls the BRLMM algorithm from the Affymetrix Power Tools (APT) for extracting genotyping calls for Affymetrix 50/100k array set. For complete details on BRLMM, refer to the Affymetrix

white paper [12]. The flow chart in Figure 26.18 reproduced from the white paper details the tasks carried out within BRLMM.

Extracting Allele Signals: The first step in analyzing the 100K arrays is the generation of signals for each SNP and each allele (A,B). Signal generation uses Quantile Normalization followed by RMA. No background correction is performed (the signal values are typically large and mileage obtained from Background Correction is expected to be limited). Quantile normalization is run on all SNPs together (i.e., when Hind and Xba arrays are analyzed together, SNPs from both are pooled together for quantile normalization.) Signal Generation involves running RMA twice, once each on the A and B alleles; these are the allele specific signals. The combined signal is the average of these two signals.

Deriving the prototypes: In parallel, BRLMM derives an initial guess for each SNPs genotype using the DM approach (with confidence threshold set to 0.17 for high stringency). It then looks across SNPs to identify cases where there are at least a certain minimum number of examples of each of the 3 genotypes according to the initial guesses. This subset of SNPs is used to estimate a prior distribution on the typical cluster centers and variance-covariance matrices. Each SNP is then visited in turn and the cluster centers and variances implied by the initial genotype guesses are combined with the prior information in an ad-hoc Bayesian procedure to derive a posterior estimate of cluster centers and variances (it is principally this step that distinguishes BRLMM from RLMM). Finally, a genotype and confidence score is assigned for each observation according to its Mahalanobis distance from the three cluster centers.

Extracting Genotype Calls: The next step is to generate Genotype Calls. BRLMM calls genotypes by a template-matching procedure comparing the transformed allele signal values observed in an experiment to the typical values (prototype) expected for each genotype. The genotype that is closest in typical value is the one that is assigned (a minimum distance classifier). The approximate confidence in that call is based on the ratio of the nearest prototype to the second nearest prototype.

Some important points concerning BRLMM are listed in Table 26.9.

Minimum Samples	BRLMM requires at least two observations of each genotype to build the prior, so the absolute minimum number of samples required is 6, though running with this small a number is not advised.
Batch Size	While more samples will generally lead to better performance, it is found that for good datasets performance reaches a plateau with as few as 50 samples, whereas for lower-quality datasets it can take as many as 100.

Table 26.9: Additional notes on BRLMM

26.7.2 Hidden Markov Model (HMM)

Computing LOH scores requires a smoothing procedure because signals and ratios are quite noisy and do not indicate clear boundaries for regions with LOH. Hidden Markov Models serve as good rounding

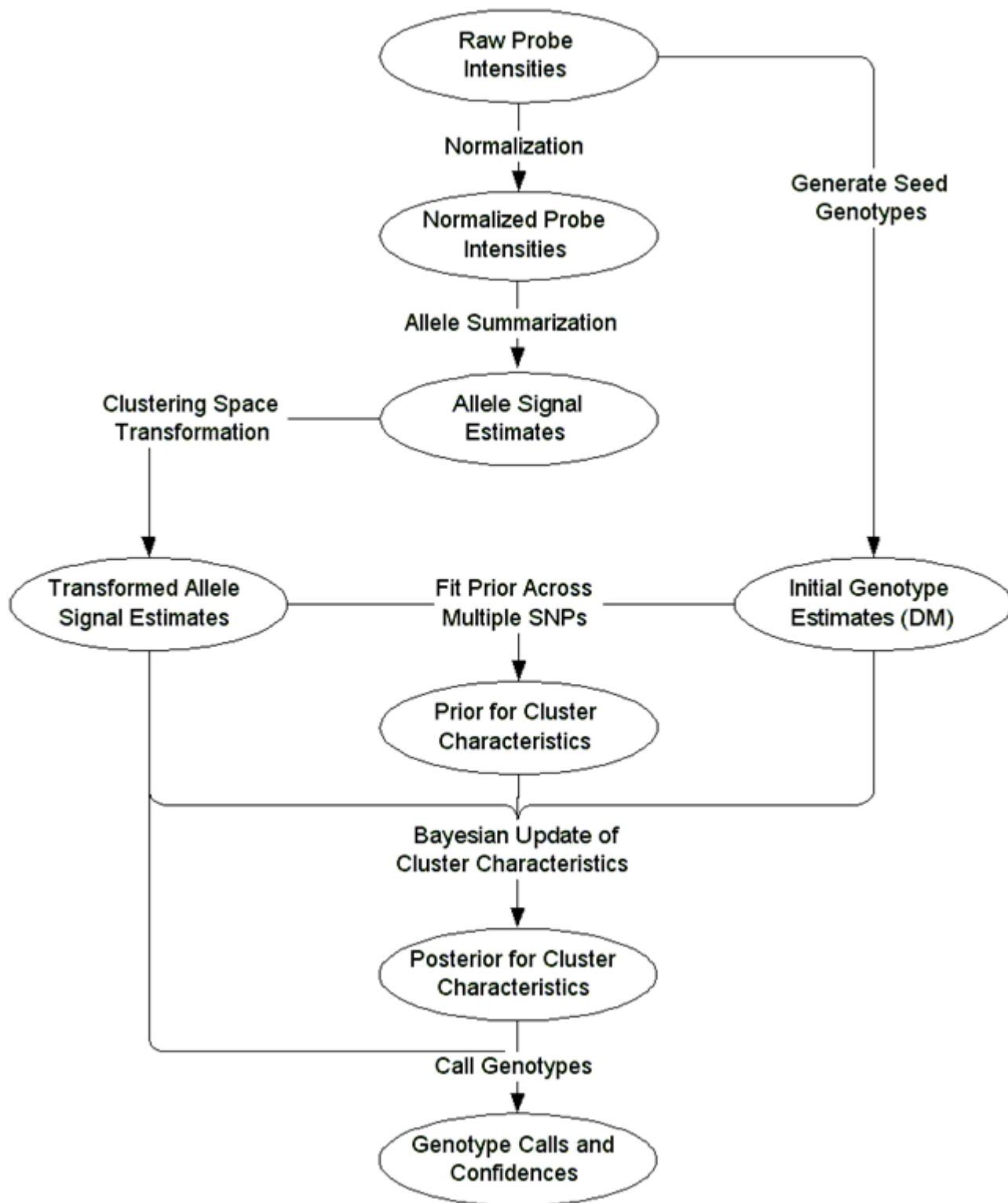


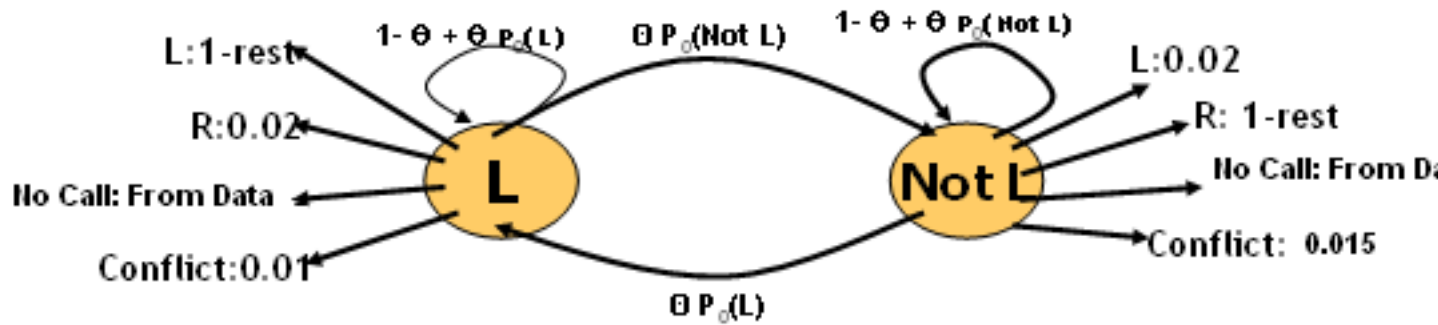
Figure 26.18: BRLMM-Flow chart

procedures in this context, because they can account for locality-based dependencies along a chromosome (i.e., if a particular SNP is in a region of LOH then the likelihood of a neighboring SNP being in a region of LOH also goes up). HMM implementation in **GeneSpring GX** is based on the reference [43].

LOH scores for analysis against a reference are generated from genotype calls using an HMM with 3 states, representing Loss of Heterozygosity (L), Retention of heterozygosity (R-HET), and Retention of Homozygosity (R-HOM), respectively. The emission probabilities at L and R-HOM are set to .99 for Homozygous and 0.01 for Heterozygous. The emission probabilities at R-Het are set to .989 for Heterozygous, 0.001 for No Calls and 0.01 for Homozygous. Transition probabilities are defined exactly as in http://galton.uchicago.edu/~loman/thesis/Thesis_double.pdf and very similar to the dChip paper <http://compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0020041> and are recapitulated in the image shown in figure Fig 26.19.

To ↓, From →	LOH	Hom-R	Het-R
LOH	$(1-\theta)+\theta P_o(L)$	$\theta P_o(L)$	$\theta P_o(L)$
Hom-R	$\theta P_o(R)P(\text{Hom})$	$[(1-\theta)+\theta P_o(R)]P_R(\text{Hom} \text{Hom})$	$[(1-\theta)+\theta P_o(R)]P_R(\text{Hom} \text{Het})$
Het-R	$\theta P_o(R)P(\text{Het})$	$[(1-\theta)+\theta P_o(R)]P_R(\text{Het} \text{Hom})$	$[(1-\theta)+\theta P_o(R)]P_R(\text{Het} \text{Het})$

Figure 26.19: Transition Probabilities for LOH analysis against Reference HMM



		Tumor			
		A	B	AB	No Call
Normal	A	N	N	C	N
	B	N	N	C	N
	AB	L	L	R	N
	No Call	N	N	R	N

Figure 26.20: The Paired Normal HMM

Here, $P_0(L) = .01$, $P_0(R) = 0.99$ and θ is set to $1 - e^{-2d}$ where d is the distance between the current and previous SNPs in units of 100MB. A higher value would increase the number of LOH regions detected but also increase false positives.

For analysis against reference, all the probabilities mentioned in this image above are computed from reference CEL files.

For paired normal analysis, a different (simpler) HMM shown in the figure fig 26.20 is used; the emission alphabet is no longer genotype calls but a Loss(L), Retention(R), Conflict(C) or Non-Informative call(N) computed from the paired samples as indicated in the figure. A smaller value would lead to fewer LOH calls.

Note that the L,C,N,R calls are not explicitly output in the spreadsheet.

Sex chromosomes are ignored for LOH computation.

26.7.3 Canary algorithm

V6 affymetrix uses CANARY (Copy Number Analysis Routine) to genotype 1320 regions of CNP (Copy Number Polymorphism). CANARY is a one dimensional Gaussian Mixture Model (GMM) to cluster samples into discrete Copy Number classes; these clusters are used by **GeneSpring GX** while assigning Copy Numbers to segments identified by CBS.

GeneSpring GX calls Affymetrix Power Tools to run Canary algorithm. Note that **GeneSpring GX** runs CANARY only during reference creation from 270 HapMap samples or user-defined custom reference samples.

Inputs:

1. CNP region file - contains CNP boundaries relative to the genome and a list of the probes used to calculate the intensity for each region
2. Raw CEL files - Probe summarization and scaling are done within Canary.
3. A 'prior' Model file which contains gaussian cluster parameters of discrete genotype classes observed in the same region. The priors are generated from 270 HapMap samples processed independently at the Broad Institute and Affymetrix. The prior model file contains the cluster mean, variance, and the expected cluster frequencies.

Outputs: Copy Number clusters for CN probes.

References:

1. <http://www.affymetrix.com/support/developer/powertools/changelog/apt-canary.html> for complete information on Canary algorithm implementation.
2. http://www.affymetrix.com/support/technical/whitepapers/canary_algorithm_whitepaper.pdf for Affymetrix White Paper on Canary Algorithm
3. Reference [9].

26.7.4 Birdseed algorithm

Birdseed is the algorithm used to extract genotype calls from Affymetrix Genome-Wide Human SNP Array 6.0, Genome-wide Human SNP array 5.0, and Human Mapping 500K Array Set. Birdseed uses a customized Expectation-Maximization (EM) algorithm to fit two dimensional Gaussians to SNP data, producing genotypes and confidence scores for every individual at every SNP. **GeneSpring GX** calls Birdseed from Affymetrix Power Tools.

GeneSpring GX runs Birdseed algorithm to create references from HapMap samples as well as to create custom references, from user inputs. Table 26.10 gives an overview.

Inputs	<ol style="list-style-type: none"> 1. Intensities from a set of CEL files for each SNP allele. 2. A 'priors' model file of gaussian clusters representing copy neutral SNPs (either from HapMap samples or from user specified reference samples).
Outputs	<ol style="list-style-type: none"> 1. Genotype calls as 0(AA), 1(AB), 2(BB) or -1(No call) at each SNP. 2. Confidence scores ranging from 0 (Least confident) to 1 (Most confident) 3. Posterior Model file containing posterior models that characterize genotype and allele specific probe responses for each SNP. This file is then used by the CBS and Fawkes algorithm.
Useful Tips	<ol style="list-style-type: none"> 1. Because it is a clustering algorithm, Birdseed should typically be run on many samples at a time (50 or more). 2. Internally, Birdseed determines the gender of each sample.
References	<ol style="list-style-type: none"> 1. http://www.affymetrix.com/support/developer/powertools/changelog/apt-probeset-genotype.html#quickstartbirdseed for more details on implementation. 2. http://www.broadinstitute.org/mpg/birdsuite/birdseed.html 3. [9] for technical information.

Table 26.10: Snap-shot of Birdseed Algorithm

Algorithm For the most part, this algorithm is a 2 dimensional GMM analogous to the 1 dimensional Canary; the two dimensions are summarized probe intensities for each of the two alleles.

Uses Expectation-Maximization algorithm as follows. The models are used for initialization. Each sample is then assigned a probability of belonging to each cluster (Estimation). Next, each cluster is redefined based off the samples that belong to it (Maximization), as well as being tethered to the expected location of the model. The expectation and maximization steps are iterated until convergence, at which point one can assign the likelihood of the model (gaussian parameters). The likelihood of the model is dependent of how well the model explains the observed as well as how the model fits certain expectations (for example, that the clusters are evenly spaced).

Birdseed chooses between models built from different initializations and between 1,2 and 3 clusters explaining the data. If the best model has fewer than 3 clusters, genotype classes corresponding to clusters not in the model are imputed to increase sensitivity to rare genotypes. The resulting 3 clusters represent the probe responses to each genotype class on the particular batch being run.

Genotyping and confidence call Let the sample i be genotyped by the cluster j (i.e, the cluster j maximised the probability when substituted by the intensities of sample i). The formulae of the weighted gaussian cluster is:

$$P\left(\frac{j}{i}\right) = \left(\frac{w_j}{2\pi \times |\sum j|^{\frac{1}{2}}}\right) e^{(-\frac{1}{2}(X_i - j^T)(\sum j^{-1}(X_i - j)))}$$

Global confidence score The relative likelihood of belonging to the next cluster

$$conf_{(global-i)} = \frac{P(j_{(second-best)}|i)}{P(j_{best}|i)}$$

Local confidence score

$$q_i = \sqrt{[(-\frac{1}{2} \times (X_i - j)^T)(\sum j^{-1}(X_i - j))]} \\ conf_{local-i} = \frac{1}{1+e^{(p-q_i)}} - \frac{1}{1+e^p}$$

where q is the number of standard deviations a sample is from its assigned cluster and p is the number of standard deviations beyond which the local confidence score quickly increases. Note that the default $p = 4.0$.

The overall confidence score Affymetrix Birdseed 'Overall confidence' score is then

$$Confidence = (c conf_{global-i} + (1 - c)(conf_{local-i}))$$

where c is the relative confidence contribution parameter, $0 \leq c \leq 1$, DEFAULT $c = 0.8$.

26.7.5 CBS for segmenting genome with respect to Copy Number

The Circular Binary Segmentation (CBS) algorithm finds the break points in a sequence of log ratios into segments with significantly different distributions. The break points are found in a recursive way till no further significant segmentation is possible.

Input to CBS are the Log ratios while the outputs are Segment Break Points and Mean Log Ratios of each segment. Table 26.11 gives an overview of the process.

Table 26.11: Snap-shot of CBS Algorithm

Smooth outliers:	The log ratios need to be smoothed by removing outliers.
Segmentation:	Finds change points in genome using a statistic to identify a ternary break as explained in the reference [16].
Validate Change Points:	Gives option to choose T test or Permutation test for validation the segmentation. Permutation test is slower and takes about 30 minutes per sample. T test is faster. Default option is T test and can be changed from the configuration options. Number of permutation or the p value cut off can also be set from here. Default number of permutations is 5000 while default p value cut-off is 0.002.
Configuration options:	Many other parameters in CBS are configurable as explained in section Configuration Options for Copy Number Analysis .

References for CBS:

1. *Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data* [14] for basic technical information on CBS.
2. *Change-point methods for the analysis of array-based DNA Copy Number Data - Adam B. Olshen* [15]
3. For latest updates on CBS algorithm, check the website quoted in this reference [16].

26.7.6 Post Processing to assign Copy Numbers to segments created by CBS

Once the segments are identified and validated by CBS, Copy Numbers have to be assigned to the segments.

- Inputs:**
- Reference clusters from Birdseed and Canary algorithms.
 - Copy Number segments from CBS.
 - Mean log ratios of segments from CBS.

- Process:**
- For each of the Copy Number clusters (0,1,2,3,4) obtained from Birdseed and Canary algorithms, there is a mean value assigned to each probe. Call this as $m_0, m_1, m_2, m_3,$ and m_4 .
 - Get the Log Mean Ratio by taking the ratio $m_0/m_2, m_1/m_2, m_2/m_2, m_3/m_2, m_4/m_2$ and then the logarithm to the base of 2 of this ratio.
 - There will be one Log Mean Ratio for each of the Copy Number values for each probe. ('0' value for Copy Number 2)
 - Get a median of the log mean ratios across all the probes for each Copy Number value to get 5 medians, M_0, M_1, M_2, M_3 and M_4 . This is the 'Median Map'. Standard deviation value is also available for each cluster.
 - For Affymetrix Mapping 100k array set, since there are no clusters (only BRLMM runs on this array), clusters generated from Genome-Wide Human SNP Array 6.0 technology of the reference are used here.
 - Take the Mean Log Ratio of each segment from CBS and check with the 'Median Map' described above
 - A fitting is done as explained below.
 - Using the standard deviation associated with each cluster, a distribution curve is created (assuming Gaussian distribution) for each cluster and the overlap point of consecutive clusters are identified.
 - The overlap point between Copy Number 0 and 1 is assigned a value of 0.5; similarly 1.5 for overlap between 1 and 2, and so on for the other clusters.
 - The mean log ratio is mapped to this distribution and values are assigned.
 - If the standard deviation is '0' for any cluster, then a fresh value is assigned by taking the mean of the standard deviation values of adjacent clusters and dividing it by the value of the Copy Number of the cluster which is in-between.

- Outputs:**
- Copy Numbers for each segment
 - Confidence:
 - For each segment, if the Mean Log Ratio corresponds to Copy Number in the range (Normal +/- 0.5), a p value of '1' is assigned.
 - For segments with Mean Log Ratio mapping to Copy Numbers out of the range 1.5 to 2.5, a T test is done against '0' and a p value is obtained.
 - Multiple testing correction is applied to this p value by multiplying it by the number of segments which are NOT of Copy Numbers 1.5 to 2.5 to get final p value.
 - Negative logarithm to the base 10 of the final p value reported as confidence.
 - In Paired Normal Analysis, for X and Y chromosomes of Normal Male samples, the confidence is set to 0 if the Copy Number is between 0.5 and 1.5.

Corrections for Sex chromosomes: 'Male' is inferred for Copy Number less than '1.5'. In Paired Normal Analysis, for males, for both X and Y chromosomes, '1' is subtracted as correction. For females, there is no correction for Chromosome X while Y chromosome is set to '0'.

26.7.7 Fawkes algorithm

Fawkes (Fast Analysis With Kopy-number et SNPs) produces SNP genotypes of the form $n.m,c$ where n is the non negative integer number of the copies of the A allele

m is the non negative number of copies of the B allele
c is the floating point confidence of this call, where 0 is most confident and 1 is least confident.

Table 26.12 gives an overview.

Inputs	<ol style="list-style-type: none"> 1. Total Copy Number 2. Gaussian cluster data from Birdseed algorithm 3. Allele intensities at each SNP
Outputs	<ol style="list-style-type: none"> 1. Allele Specific Copy Number 2. Confidence associated with the above
Notes	The Allele Specific Copy Number generated by Fawkes is the input for the LOH analysis using HMM. For Paired Normal analysis, the input to LOH are the genotype calls generated by Fawkes by setting the Copy Number of all 'Normal' samples to '2'.
References	<ol style="list-style-type: none"> 1. [9] for technical information.

Table 26.12: Snap-shot of Fawkes Algorithm

Algorithm Briefly, the gaussian cluster outputs of Birdseed (associated with copy neutral calls) are extrapolated to generate gaussian clusters for Copy Number 0, 1, 3, 4 ,etc. The intensities of the SNP (X_A, X_B) are substituted into each cluster and the cluster with the maximum probability decides the genotype call.

- For example, extrapolation of clusters for Copy Number 3 would imply generating 4 clusters, A3B0, A2B1, A1B2, A0B3.
- The 'Means; of Copy Numbers 0, 1, 2 are extracted from the Birdseed clusters and both the 'Means' and 'Variances' for all other Copy Numbers are linearly extrapolated as a function of 'Means'.

The underlying A and B allele frequencies w_a and w_b are estimated using the frequencies of the AA, AB and BB clusters.

$$\begin{aligned}
 w_a &= 1 - w_b \\
 &= \frac{2w_{AA} + w_{AB}}{2} w_j = w_a^N * w_b^M * \frac{N + M - 1!}{N! * M!}
 \end{aligned}$$

For genotype cluster j, with N copies of A allele and M copies of B allele.

Confidence This is exactly as per [Birdseed Overall Confidence Score](#)

26.8 Tutorials for Copy Number Analysis

Tutorial for Copy Number Analysis is available at <http://www.chem.agilent.com/en-US/Products/software/lifesciencesinformatics/genespringgx/pages/gp34528.aspx>.

Chapter 27

Association Analysis

This chapter provides instructions to run an Association experiment using **GeneSpring GX** .

Tutorials: You can also refer to the [tutorials](#) page on Association experiment.

27.1 Introduction

GeneSpring GX provides a toolkit for Genome-Wide Association Analysis (GWAS). The focus of the Association Analysis lies on the analysis of SNP and phenotype data for studies that can involve thousands of unrelated samples genotyped at hundreds of thousands of SNPs.

It includes a variety of features:

- Quality Control (QC) methods to identify anomalous samples which would confound the analysis.
- Filters to remove undesirable SNPs from the analysis.
- Rigorous Principal Components Analysis (PCA)-based methods for identifying and removing sources of stratification in the experiment.
- Statistical tests to identify SNPs that are associated with disease incidence, and methods to eliminate **false alarm** results.
- Haplotype inference and analysis methods, which can detect more subtle disease-causing groups of SNPs.
- A wide range of publication-quality intuitive visualization options, which include LD Plot, Haplo block view, etc.

- Full integration into other features of GeneSpring GX, such as Pathway Analysis and Genome Browser.

In **GeneSpring GX** you can create Association Analysis experiments for the Affymetrix and Illumina chip technologies, as well as import genotype data in a text file.

Note: Illumina Association Analysis experiments do not require any algorithm to fetch the genotype calls.

27.2 Technology

GeneSpring GX supports the following technologies:

Technology	Algorithm
Affymetrix Mapping 50K Xba240	BRLMM
Affymetrix Mapping 50K Hind240	BRLMM
50K Xba240 and 50K Hind240	BRLMM
Affymetrix Mapping 250K Nsp	Birdseed
Affymetrix Mapping 250K Sty	Birdseed
250K Nsp and 250K Sty	Birdseed
Affymetrix GenomeWide SNP5	Birdseed
Affymetrix GenomeWide SNP6	Birdseed
Illumina (all)	None; genotype calls are generated in GenomeStudio.
Custom (Illumina Format)	None; genotype calls are present in the input file.

Table 27.1: Technologies and Genotype Call Algorithms for Association Analysis Experiments

Notes:

- You need at least six distinct CEL files for Hind or Xba experiments.
- You need at least six distinct file pairs for combined Hind and Xba experiments.

27.3 Experiment Creation

GeneSpring GX allows you to create Affymetrix, Illumina, and Custom Association experiments. You can run Affymetrix Association experiments on CEL files (refer to [Technology](#) section for information on the technologies and Genotype Call Algorithms).

You can also run Illumina Association experiments on .txt files containing genotype information created using Illumina's GenomeStudio. You can import genotype information as a .txt file from Genome Studio using a plug-in (refer to section 27.3.1).

You can also run Custom Association experiments on .txt files containing genotype information in the Illumina GenomeStudio format (refer to section 27.3.1).

27.3.1 Illumina Association Analysis Experiment

You can create an Illumina Association Experiment from the Project menu (*Project menu* → *New Experiment* → *Illumina Association Analysis* as Experiment Type in the *Experiment Description* dialog), or from an existing Experiment (*Context menu* → *New Experiment*).

The Illumina Association Analysis Experiment wizard has one step.

On the Wizard page, click on Choose Files or Samples button to open the file browser or the Sample/RawFile Search Wizard. You can select the files and samples using the browser and the wizard respectively.

The Sample/RawFile Search wizard has two steps.

Step 1:

Allows you to Add, Remove, and Combine Search Conditions (by "OR" or "AND"). You can select "Show User Attributes" check box to open the search results with the user attributes, and enter an appropriate value for the "Max Results per page" (default value: 100).

Step 2:

Launches the search results of samples (select and double click to launch the Sample Inspector). Select the Samples you want to use in the Experiment, and click "Finish".

- Click on "Reorder" button to change the ordering of the samples.
- Click "Finish".

Illumina File Format

This section provides an overview of the File (.txt) Format for Illumina Association Experiments.

The .txt files have five technology columns:

- **Name:** Provides the name of the probe.
- **Chr:** Provides the Chromosome number.
- **Position:** Provides the physical position of the probe.
- **HW Equil:** Provides the Hardy–Weinberg (HW) Equilibrium score for the SNP.
- **CNV Region:** Provides information on any nonpolymorphic probes falling in the region. This column is automatically populated with information, and may not be current because the number of known Copy Number Variant(CNV) regions is constantly changing. This column is for informational purposes only.

The.txt files also have six sample-specific columns:

- **GType-Sample:** Provides the genotype of the subject SNP for the sample.
- **Score:** Provides Genotype Call score of the subject SNP for the sample.
- **B Allele Frequency:** Provides the theta value for an SNP, corrected for cluster position. Cluster positions are generated from a large set of normal individuals.
- **Log R Ratio:** Provides the log (base 2) ratio of the normalized R value for the SNP divided by the expected normalized R value.
- **CNV Value:** Provides an estimate of Copy Number at individual locus.
- **CNV Confidence:** Provides the level of confidence that the CNV value is correct, based on the CNV algorithm used.

Importing Illumina Files

You can export genotype information as a .txt file from GenomeStudio. This would require you to register the GeneSpring plug-in in the GenomeStudio. The following instructions will guide you in exporting genotype information from GenomeStudio:

1. Copy the plug-in from `INSTALLDIR\app\Illumina\GX.Genotyping.Export.dll` to `Genomestudio\modules\BSGT\ReportPlugins\`.
2. Restart GenomeStudio.
3. Double click on a Project from the **Recent Projects** pane in the GenomeStudio **start page**. You can also create a project from the **New Projects** pane, running the project creation wizards. You can use these wizards to create different types of projects, e.g., Genotyping, Gene Expression, etc. (refer to GenomeStudio User Manual for more information).

<p>Note: The GeneSpring plug-in is only meant for Genotyping Projects.</p>
--

4. Select the **Reports** option from the the **Analysis** menu.
5. Click **Report Wizard** option from the **Report** sub menu. This launches the **Report Type** wizard page.
6. Select **Custom Report** radio button, and **GeneSpring Exporter 1.0 from Strand Life Sciences** from the drop-down list, and click **Next**. This launches the second page of the **Sample Groups** wizard.
7. Select the sample groups you want to include in the report, and click **Next**. This launches the **Destination** wizard page.
8. Provide an **Output Path** (you can use the **Browse** button), and **Report Name**, and click **Finish**. This launches a **Progress bar** with information about the incremental progress in generating the Report. After the Report is written completely, you will see a message with an option to view the Report.

Custom Technology

You can create a technology name from the "New Illumina Association Analysis" wizard.

- Select **EnterNew** option from the **Select a technology name** drop-down box.
- Enter a name relevant for the experiment.

If you want to create an experiment with a Custom Technology, you can create a sample .txt file and run an "Illumina Association Analysis" Experiment. All technology and sample-specific columns should be present in this sample file (refer to section 27.3.1). The **GType-Sample** column is the only one which will be used hereafter in this experiment.

27.3.2 Affymetrix Association Analysis Experiment

You can create an Affymetrix Association Experiment from the **Project** menu (*Project menu* → *New Experiment* → *Affymetrix Association Analysis* as *Experiment Type* in the *Experiment Description* dialog), or from an existing Experiment (*Context menu* → *New Experiment*).

Refer to [Creating a Copy Number Experiment](#) section for the steps involved in the process.

27.4 Quality Control

In **GeneSpring GX** you can filter samples for missing genotype calls, and run the birdseed report and EIGENSTRAT filter on samples.

- Filter samples by Missing values: Removes samples for which too many of the SNPs did not return a valid genotype.
- Run "Birdseed report" on samples: Displays a summary of results if the Birdseed genotype call algorithm is being used (for certain Affymetrix technologies; refer to Table 27.1).
- EIGENSTRAT filter on samples: Identifies samples that are contributing heavily to stratification in the experiment.

27.4.1 Filter Samples by Missing Values

You can identify samples that have a higher proportion of missing genotype calls, across SNPs. A high proportion of missed calls for a given sample can indicate an anomalous or inferior genotyping process for the sample, and can make later analyses less valuable. It is therefore customary to exclude these samples in further analysis. You can set a cutoff for missingness (default value= 0.1), and the samples that have missingness rates greater than the cutoff value are removed.

Notes:

- You should choose an appropriate cutoff value to filter out samples that have higher missing genotype calls, and are less relevant for the experiment.
- You can remove samples using the Add/Remove Samples button, and add the samples back if required.

Input	<ul style="list-style-type: none"> • Active entitylist • Cutoff for missingness (default value= 0.1)
Process	<ul style="list-style-type: none"> • Calculates the proportion of missing genotype call, combined across SNPs for each sample. • Filters out samples that have the proportion of missing genotype calls greater than the threshold.
Output	List of samples to be kept.

Table 27.2: Summary of Steps: Filter Samples by Missing Values

Note: It is not advisable to apply this filter with settings that result in a very unequal number of cases and controls. For case-control association studies, best results are achieved with an approximately equal number of cases and controls.

27.4.2 Birdseed Report

You can launch a summary of the Birdseed algorithm's results if the experiment technology uses the algorithm (refer to Table 27.1). It provides a quick glance at the input CEL files, at a more detailed level than just the genotype calls, which are used exclusively by the rest of the workflow. Refer to 27.3 for more information on each column.

Column	Description
cel files	CEL file name.
computed gender	Estimated gender.
call rate	Genotype call rate at the default or user-specified threshold.
het rate	Percentage of "AB" genotype call (i.e., the heterozygosity).
hom rate	Percentage of "AA" or "BB" genotype call (i.e. the homozygosity).
cluster distance mean	Average distance to the cluster center for the genotype.
cluster distance stdev	Standard deviation of the distance to the cluster center for the genotype.
raw intensity mean	Average of the raw PM probe intensities.
raw intensity stdev	Standard deviation of the raw PM probe intensities.
allele summarization mean	Average of the allele signal estimates (log2 scale).
allele summarization stdev	Standard deviation of the allele signal estimates (log2 scale).
allele deviation mean	Average of the absolute difference between the log (base 2) transformed allele signal estimate and its median across all chips.
allele deviation stdev	Standard deviation of the absolute difference between the log transformed (base 2) allele signal estimate and its median across all chips.
allele mad residuals mean	Average of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
allele mad residuals stdev	Standard deviation of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
em-cluster-chrX-het-contrast gender	Gender estimate based on estimated heterozygosity on chrX.
em-cluster-chrX-het-contrast gender chrX het rate	Estimated heterozygosity on chrX.
cn-probe-chrXY-ratio gender meanX	Average intensity of chrX CN probes.
cn-probe-chrXY-ratio gender meanY	Average intensity of chrY CN probes.
cn-probe-chrXY-ratio gender ratio	Ratio of average chrY CN probe intensity to average chrX CN probe intensity.
cn-probe-chrXY-ratio gender	Gender estimate based on ratio of chrY to chrX average CN probe intensities.

Table 27.3: Birdseed Report

Refer to the section on [Birdseed Algorithm](#) for more information on the Algorithm.

27.4.3 EIGENSTRAT Filter on Samples

The EIGENSTRAT filter allows you to remove the samples contributing heavily to the existing stratification [49] in the experiment.

Stratification is an undesirable phenomenon that occurs when the samples used in the experiment can be bracketed into distinct groups that show different characteristics with respect to the condition (e.g., disease) being studied.

Suppose all cases were race A and all controls were race B, and race A was known to have a far greater susceptibility to the disease than race B. This situation would confound the analysis, because any association could simply be attributed to the race differences between cases and controls. We say in this case that the samples exhibit high stratification, and the racial difference is a source of stratification.

A well-designed experiment would ideally have little to no stratification, but sources of stratification are often almost impossible to detect or compensate for. This step uses PCA to identify samples which contribute disproportionately to the experiment's stratification. If they are found, it is often advisable to exclude these samples from further analysis to drastically reduce the confounding effects of stratification.

The filter identifies the principal components that contribute most to the covariance between the samples. Those that contribute significantly more than expected (as measured by a test called the TW test) are interpreted as **axes of stratification**. The filter shows the z-score for each sample along the axes of stratification. You can filter out the samples that have z-score magnitudes greater than a certain cutoff value. The following steps will guide you to filter out samples that contribute heavily to the existing stratification (refer to [PCA](#) section).

The following steps will guide you to filter out samples that contribute heavily to the existing stratification:

1. Check if PC_1 is significant (i.e., p-value less than the cutoff). Click the **Change Cutoff** button to adjust the TW p-value cutoff value (refer to the [notes on change cutoff options](#)). There is no stratification if PC_1 is not significant.
2. If PC_1 is significant, check PC_2, PC_3, PC_4 , etc., in sequence for significance. Continue till you come across an insignificant PC.
3. If the first k PCs (i.e., $PC_1, PC_2, \dots, PC_{k-1}, PC_k$) in step 2 are significant, but $(k+1)^{th}$ is not, then configure the EIGENSTRAT Filter parameters.
 - Adjust the **Top PCs to display** slider to launch the plot with only the first k PCs.
 - Adjust the **TW p-value cutoff** slider correctly as desired.

- In the **z-scores** tab of the bottom pane, the samples with z-scores whose magnitudes are greater than the z-score cutoff along any significant PC (in this case, the first k PCs) are highlighted. The z-score cutoff can be changed using the "Change cutoff" button.

Note: A higher z-score implies a greater contribution to the existing stratification.

4. If required, you can re-run the PCA, and then repeat steps 1–3.

Restoring removed samples:
Run one of the QC filters (e.g., filter samples by missingness), and use the "add/remove samples" button and then select the samples you want to restore.

Input	<ul style="list-style-type: none"> • Active entitylist • TW P-value cutoff for the Principal Components • Z-score cutoff for the samples (default value: maximum z-score across all samples and PCs) • Number of top PCs to display (default value: p-2; where p is the number of samples)
Process	<ul style="list-style-type: none"> • Identifies the directions of stratification, and groups the samples. • Finds samples that contribute heavily to the existing stratification (absolute value of z-score greater than the cutoff), and filters them out.
Output	List of samples to be kept.

Table 27.4: Summary of Steps: EIGENSTRAT Filter

EIGENSTRAT notes on minimum number of samples/SNPs for successful operation

- A very low number of samples (less than 30, in many situations) may lead to inferior performance. It is advisable to use Genomic Control in such situations.
- Higher stratification requires less SNPs for EIGENSTRAT to detect and correct for it.
- 100K SNPs are almost always enough, 20k are enough for most studies, and less than 5k rarely works.

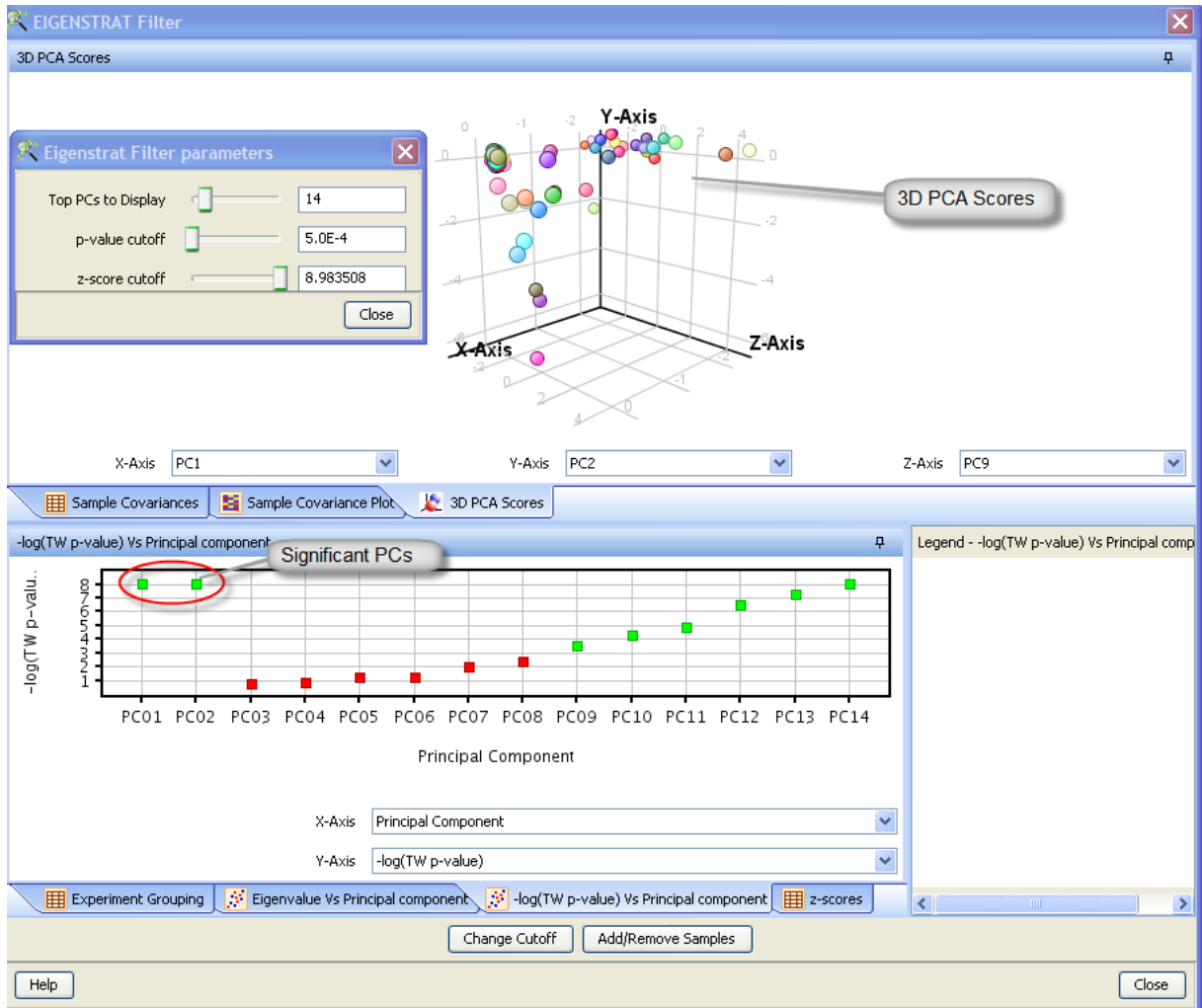


Figure 27.1: EIGENSTRAT Filter

EIGENSTRAT Filter dialog

The result screen has two panes (Figure 27.1). The top pane has three tab options, viz., Sample Covariances, Sample Covariance Plot, and 3D PCA Scores (default view). The bottom pane has four tab options, viz., Experiment Grouping, Eigenvalue vs PC plot, $-\log(\text{TW p-value})$ vs PC plot (default view), and Z-scores.

Notes on change cutoff options:

- You can move the *Top PCs to Display* slider to adjust the number of PCs to be displayed in the plots. (default value: p-2; where p is the total number of samples).
- You can move the *TW p-value cutoff* slider to adjust the threshold for PC significance. If the TW p-values for all the PCs are above the cutoff, then the **3D PCA scores** and **z-scores** tabs are removed from the screen, since there will be no **significant PC** over which z-scores could be calculated. Also, the z-score cutoff automatically changes to -1 to further reflect this.
- You can move the *z-score cutoff* slider to adjust the threshold for absolute value of z-score. This cutoff is set to -1 if TW p-values for all the PCs are above the cutoff (see above).

Sample Covariance: Contains the covariance matrix ($X^T X$) of normalized genotype calls.

The covariance matrix is the product of the normalized genotype calls matrix (X) for n SNPs across p samples and its transpose matrix (X^T).

Sample Covariance Plot: Launches a heatmap of the covariance matrix. You can configure the color range from the Properties option on the context menu.

3D PCA Scores: Plots each sample as per the z-scores along the three chosen PCs.

Notes:

- You can select the PCs from the X-, Y-, and Z-axis drop-down boxes.
- You can select any sample from the plot canvas, and then filter it out using the **Add/Remove Samples** option.
- You can zoom into the plot canvas using Ctrl + left-click + mouse hover from bottom to top. Press Ctrl + left-click + mouse hover from top to bottom to zoom out.
- You can rotate the plot canvas using Ctrl + left-click + mouse hover. The direction of mouse hover determines the axis, and rotation (clockwise/anti-clockwise).

Experiment Grouping: Shows the experiment grouping details.

Eigenvalue vs PC plot: Plots the eigenvalue for each PC.

$-\log_{10}(\text{TWp-value})$ vs PC plot: Plots the transformed Tracy–Widom p-value for each PC.

z-score: Provides the z-score for each sample across all the significant PCs.

Notes:

- The EIGENSTRAT filter follows the same procedure as EIGENSTRAT correction to select significant PCs. Removes samples according to cutoffs. You should keep the z-score cutoff (suggested default: 6) such that only a few samples are filtered out.

You should remove only a few samples in this step, and examine all the samples that could be removed on a case-by-case basis in the context of the experiment or how data collection was performed.

- When an SNP has the same genotype for all the samples EIGENSTRAT ignores it completely.

27.5 Filters

In **GeneSpring GX** you can Filter SNPs by Missing Value, Identify SNPs with Differential Missingness, and Filter SNPs by HWE p-value and MAF.

GeneSpring GX offers several filters for removing SNPs which are undesirable in some way from the analysis:

Filter SNPs by Missing Value:

Allows you to remove SNPs which show 'missing' genotype calls for an unacceptably high proportion of samples.

Identify SNPs with Differential Missingness:

Identifies SNPs for which missingness patterns seem to significantly differ between cases and controls.

Filter SNPs by HWE p-value:

Allows removal of SNPs whose genotype frequencies deviate too much from Hardy-Weinberg Equilibrium, a standard genotype frequency baseline.

Filter SNPs by MAF:

Allows removal of SNPs for which the genotype calls (across samples) are dominated too much by one allele to the exclusion of the other.

27.5.1 Filter SNPs by Missing Value

If an SNP has too many missing calls (i.e., the proportion is above a set cutoff, default value= 0.1), it is not worth analyzing. Here, you can filter out such SNPs. Once you set a cutoff value (default value= 0.1), the SNPs with missingness rates greater than the cutoff are identified and filtered out.

Input	<ul style="list-style-type: none"> • Entity List • Cutoff for missingness (default value= 0.1) Note: You should choose an appropriate cutoff value to filter out SNPs that have higher missing genotype calls, and are less relevant for the experiment.
Process	<ul style="list-style-type: none"> • Calculates the proportion of missing genotypes (combined across samples). • Counts the number of missing genotype calls added over all samples. Note: The missing calls occur in pairs, i.e., either both or none of the alleles for an SNP are present. Only one cannot be missing. The number of missing calls is an even number. • Divides the missing calls by twice the number of samples, which is the total number of allele calls for the SNP. The result is the missing call proportion. • Filters out the SNP if the proportion is above the cutoff.
Output	List of SNPs to be kept.

Table 27.5: Summary of Steps: Filter SNPs by Missing Value

27.5.2 Identify SNPs with Differential Missingness

In **GeneSpring GX** you can filter out SNPs that show a strong correlation between missingness and case–control status. A χ^2 (chi-square) test quantifies the significance of this correlation. This filter displays SNPs with a p-value less than a specified cutoff (default value= 0.05).

For a given SNP, if 95% of the missing genotypes occurred among the cases and only 5% among the controls, there could be good reason to suspect a correlation. Such SNPs could show a disease-dependent characteristic which interfered with the genotyping process, making any association results (which would ignore the missing calls) suspect.

The significance of any correlation is measured using a χ^2 test (refer to Table 27.6).

Input	<ul style="list-style-type: none"> • Genotype information for all SNPs, across samples • Disease Status of each sample • P-value cutoff (default value= 0.05)
Process	<ul style="list-style-type: none"> • Calculates the missing genotypes for each SNP, combined across samples. • Counts the number of missing genotype calls added over all samples. • Calculates the χ^2 statistic for each SNP. • Finds the p-values, given the χ^2 statistic is distributed as chi-square with one degree of freedom. • Filters out the SNP if the p-value is less than the cutoff.
Output	List of SNPs to be kept.

Table 27.6: Summary of Steps: Filter SNPs by Differential Missingness

	Missing	Non-Missing	Total
Cases	N_{MC}	N_{PC}	N_C
Controls	N_{MCcont}	N_{PCcont}	N_{Cont}
Total	N_M	N_P	N

Table 27.7: Contingency Table for Differential Missingness

Calculation

- Suppose the missing and non-missing genotype call counts for Cases and Controls are N_{MC} , N_{PC} , N_{MCcont} , and N_{PCcont} respectively.
- The expected occurrences of missing and non-missing genotype calls are given by the relations:
 $E_{MC} = (N_M \times N_C)/N$, $E_{PC} = (N_P \times N_C)/N$, $E_{MCcont} = (N_M \times N_{Cont})/N$, and $E_{PCcont} = (N_P \times N_{Cont})/N$.
- $\chi_{MC}^2 = (N_{MC} - E_{MC})^2/E_{MC}$. Similarly, χ_{MCcont}^2 , χ_{PC}^2 , and χ_{PCcont}^2 are calculated.
- $\chi^2 = \chi_{MC}^2 + \chi_{MCcont}^2 + \chi_{PC}^2 + \chi_{PCcont}^2$.

Note: The Identify SNPs with Differential Missingness filter establishes correlation between missingness and disease status; whereas Filter SNPs by Missingness calculates the aggregate missingness rate for each SNP, ignoring the disease status.

27.5.3 Filter SNPs by HWE p-value

In **GeneSpring GX** you can filter out SNPs that show a strong deviation from Hardy-Weinberg Equilibrium (HWE). A chi-square (χ^2) test quantifies the significance of this deviation. This filter removes SNPs with a p-value less than the cutoff (default value: 0.001). You can configure the p-value cutoff from the **Change cutoff** button.

HWE is a standard biological model concerning genotype frequencies that is used as a baseline for many tests. It gives the genotype frequencies for a particular SNP depending on the allele frequencies, assuming that the alleles for any given next-generation individual are chosen independently and randomly. Many tests assume HWE or perform best when it is followed, and SNPs which show large deviations from it are often pathological or anomalous. Most SNPs in large experiments will not deviate significantly from HWE, and any SNP which does is generally considered unsuitable for further analysis. In this step, you can filter out such SNPs.

A χ^2 test is used to measure the significance of any deviation. The details are below.

Note: When an SNP has the same genotype for all the samples HWE filter automatically sets the p-value to 0; the SNP clearly violates HWE and should always be filtered out.

Input	<ul style="list-style-type: none"> • Genotype information for all SNPs, across samples • p-value cutoff (default value= 0.0010)
Process	<ul style="list-style-type: none"> • Separately counts the number of occurrences of the genotypes AA, AB, and BB, across all the samples. • Calculates the χ^2 test statistic for each SNP. • Finds the p-values (χ^2 distribution with one degree of freedom). • Filters out the SNP if the p-value is below the cutoff.
Output	List of SNPs to be kept.

Table 27.8: Summary of Steps: Filter SNPs by HWE p-value

Calculation

- Suppose the three genotype call counts are N_{AA} , N_{AB} , and N_{BB} respectively.
- Let,

$$f_A = \frac{2N_{AA} + N_{AB}}{2(N_{AA} + N_{AB} + N_{BB})}.$$

- Then the expected occurrences of the each of the three genotypes are given by the relations: $E_{AA} = f_A^2(N_{AA} + N_{AB} + N_{BB})$, $E_{AB} = 2f_A(1 - f_A)(N_{AA} + N_{AB} + N_{BB})$, and $E_{BB} = (1 - f_A)^2(N_{AA} + N_{AB} + N_{BB})$.
-

$$\chi_{AA}^2 = \frac{(N_{AA} - E_{AA})^2}{E_{AA}}.$$

Similarly, χ_{AB}^2 and χ_{BB}^2 are calculated.

-

$$\chi^2 = \chi_{AA}^2 + \chi_{AB}^2 + \chi_{BB}^2.$$

27.5.4 Filter SNPs by MAF

The MAF (Minor Allele Frequency) filter removes the SNPs with minor allele frequencies less than a specified cutoff value (default value= 0.01).

For a given biallelic SNP, it is desirable not to have a very large imbalance between the frequencies of the two alleles. If for instance the lower of the two frequencies (i.e., the minor allele frequency) is too low, many tests may give untrustworthy results. Note that we use no external definitions to select the minor allele, instead defining it as the allele with the lower frequency using only the data in the experiment (ignoring missing values).

Calculation:

- The sum of the two counts is an even number because the missing calls occur in pairs.
- The frequency (f_A) of A is $N_A/(N_A + N_B)$, and B (f_B) is $N_B/(N_A + N_B)$; where N_A and N_B are the two counts.
- $\text{Min}(f_A, f_B)$ is the MAF.

Input	<ul style="list-style-type: none"> • Genotype information for all SNPs, across samples • MAF cutoff (default value= 0.01)
Process	<ul style="list-style-type: none"> • Calculates the MAF across all the samples. • Counts the occurrences of A and B separately across all the samples; A and B are the alleles for the SNP. • Filters out SNPs with MAF less than the cutoff.
Output	List of SNPs to be kept.

Table 27.9: Summary of Steps: Filter SNPs by MAF

Recommendations:

- You should use the MAF filter in every run of the tool.
- χ^2 tests cannot filter out the SNPs with MAF 0%, and instead set their p-values to 1.
- EIGENSTRAT filter doesn't remove SNPs with MAF 0%.

27.6 Analysis

In **GeneSpring GX** you can perform EIGENSTRAT Correction on Samples, Statistical Analysis, SNP Tagging, SNP Regression, Haplotype Trend Regression, and LD Analysis.

27.6.1 EIGENSTRAT Correction on Samples

EIGENSTRAT Correction on Samples allows you to correct for stratification (refer to [EIGENSTRAT Filter on Samples](#) section) in the experiment.

The method is very similar to that used in [EIGENSTRAT Filter on Samples](#) section; the matrix of genotype calls (dimension: number of SNPs \times number of samples) is subjected to PCA (similar to [EIGENSTRAT Filter on Samples](#) [49], [48]). These are found using the same method as in section 27.4.3, but are then used to modify the genotype data. The modified data will contain all salient features of the original genotype data, except that it will not show significant stratification. It can therefore be used in later analysis steps without fear of stratification-related confounding effects.

<p>Notes:</p> <ul style="list-style-type: none"> • The corrected data generated here cannot be used in some of the tests offered under the Association Analysis workflow. This is because the genotype calls have been changed to decimal numbers and therefore cannot be treated as conventional genotypes in some ways. • The phenotypes are also modified by this method, so that they become floating-point values, like the modified genotypes.

Input	<ul style="list-style-type: none"> • Active entity list • Mode of Inheritance (default option: Additive; refer to section 27.6.1) for details
Process	<ul style="list-style-type: none"> • Identifies the principal axes of variation for the data. • Corrects the data to reduce the variations along the principal axes.
Output	Corrected genotype information for all the SNPs, across samples.

Table 27.10: Summary of Steps: EIGENSTRAT Correction on Samples

The following steps will guide you in using the EIGENSTRAT Correction on Samples:

Input Parameters (Step 1) Select the Mode of Inheritance (default: Additive; refer to section 27.6.1) for EIGENSTRAT Correction.

- You can select any of the three modes, viz., Additive, Recessive, and Dominant from the drop-down box.
- You can add a new mode, and then enter the weight factors for each of the genotype calls, viz., AA, AB, and BB. (*Tools* → *Options* → *Miscellaneous* → *Mode of Inheritance*; refer to section 27.6.1)

PC Selection (Step 2) The screen has two panes (Figure 27.2). The top pane has two options, viz., Eigenvalue vs PC and TW p-value vs PC Plots (default view).

The bottom pane has three options, viz., Sample Covariances, Sample Covariance Plot, and Experiment Grouping (default view).

You can select the significant PCs (refer to EIGENSTRAT Filter section) from the active view on the top pane, and then click Next. When selecting significant PCs, it is sometimes not absolutely clear how many to select. The following rules of thumb can be used:

- Never consider more than about 10–15 PCs significant.

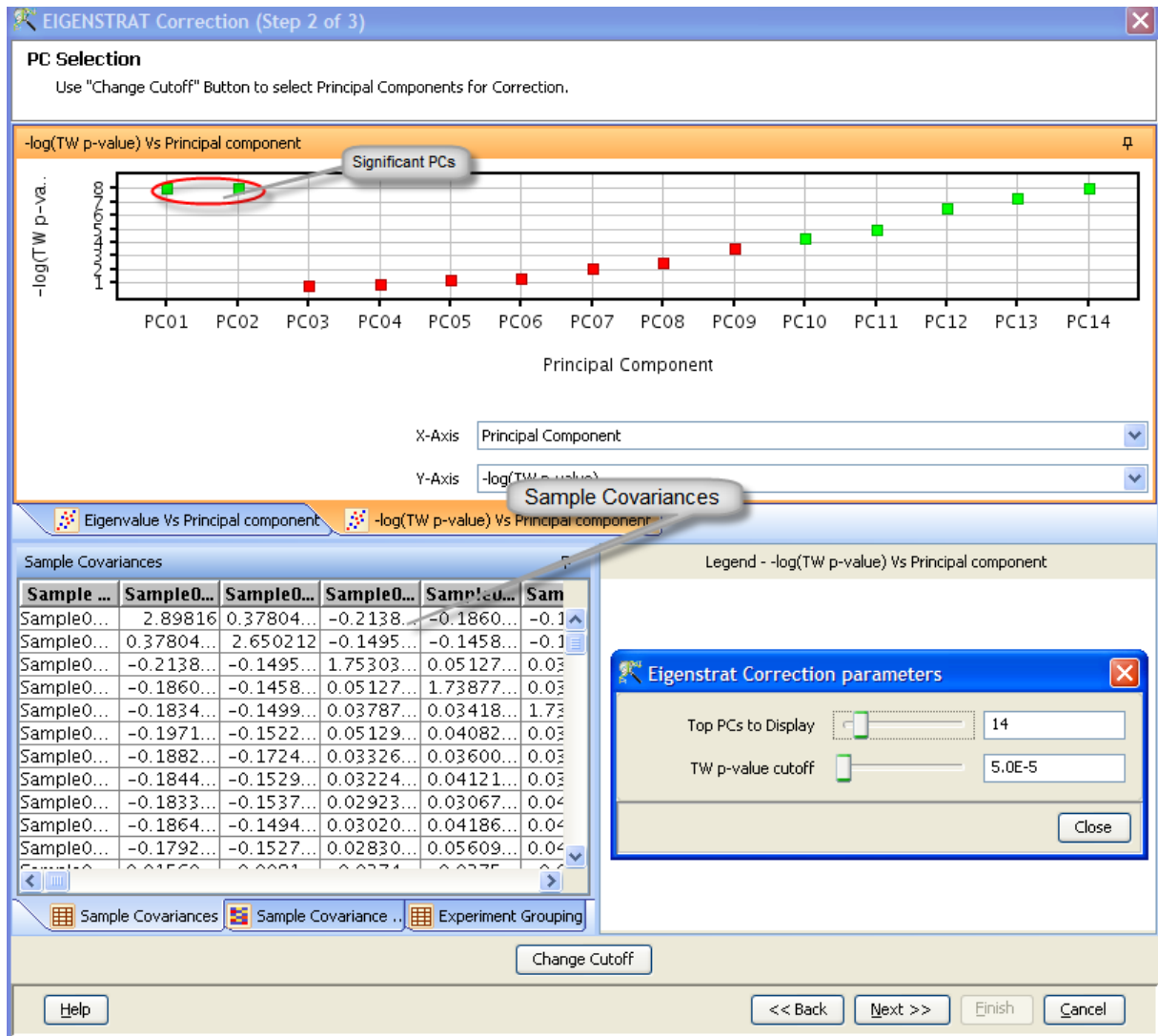


Figure 27.2: EIGENSTRAT Correction View

- When a first group of few PCs have p-values far below the next few PCs, even if the latter PCs also have p-values below the cutoff, think about only declaring the first group significant.

Note: The EIGENSTRAT filter follows the same procedure as EIGENSTRAT correction to select significant PCs.

Eigenvalue vs PC plot:

Plots the eigenvalue for each PC.

$-\log_{10}(\text{TWp} - \text{value})$ vs PC plot:

Plots the transformed Tracy–Widom p-value for each PC.

Sample Covariances:

Contains the covariance matrix ($X^T X$) of normalized genotype calls.

The covariance matrix is the product of the normalized genotype calls matrix (X) for n SNPs across p samples and its transpose matrix (X^T).

Sample Covariance plot:

Launches a heatmap of the covariance matrix. You can configure the color range using *Context (right-click)* menu \rightarrow *Properties*.

Experiment Grouping:

Shows the experiment grouping details.

Notes on change cutoff options:

- You can move the *Top PCs to Display* slider to adjust the number of PCs to be displayed in the plots. (default value: $p-2$; where p is the total number of samples).
- You can move the *TW p-value cutoff* slider to adjust the threshold for PC significance.

Results (Step 3) The window has two panes (Figure 27.3); the top pane launches the "Corrected Dataset" of genotype calls, and the bottom panes display SNP-wise and Sample-wise variation. These two are summary statistics measuring the magnitude of the correction on the data, viewed SNP-wise and sample-wise respectively.

Corrected Dataset	Provides the genotype calls matrix with the corrected data.
Sample-wise Variation	Calculates the differences between the uncorrected and corrected datasets for each sample, and then quantifies the extent of correction. This is done by taking the Euclidean norm of the difference between the uncorrected and corrected vectors for each sample.
SNP-wise Variation	Calculates the differences between the uncorrected and corrected datasets for each SNP, and then quantifies the extent of correction. This is done by taking the Euclidean norm of the difference between the uncorrected and corrected vectors for each SNP.

Table 27.11: EIGENSTRAT Correction Result Screen

Calculation:

- Let $G_{1k}, G_{2k}, \dots, G_{nk}$ be the genotype calls for n SNPs in the k^{th} sample.
- Let $C_{1k}, C_{2k}, \dots, C_{nk}$ be the corrected genotype calls for n SNPs in the k^{th} sample.
- Then, (Sample-wise) variation for the k^{th} sample = $\sqrt{\sum_{i=1}^n (C_{ik} - G_{ik})^2}$.
- Similarly, (SNP-wise) variation for the j^{th} SNP = $\sqrt{\sum_{i=1}^p (C_{ji} - G_{ji})^2}$.

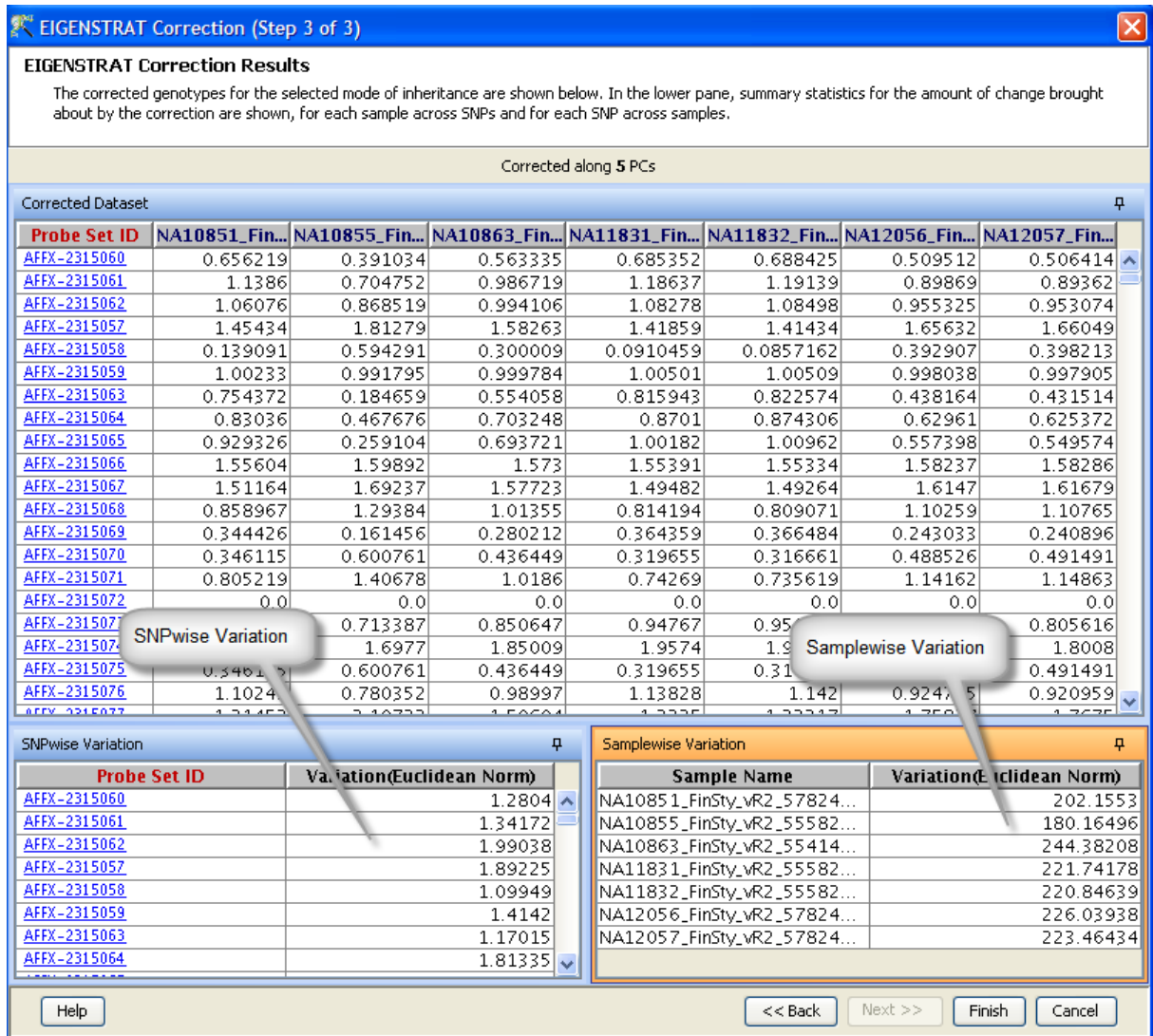


Figure 27.3: EIGENSTRAT Correction Results

EIGENSTRAT notes on minimum number of samples/SNPs for successful operation

- A very low number of samples (less than 100) may lead to inferior performance.
- Higher stratification requires less SNPs for EIGENSTRAT to detect and correct for it.
- 100K SNPs are almost always enough, 20k are enough for most studies, and less than 5k rarely works.

Mode of Inheritance

The mode of inheritance for the condition being studied is the relationship between the expressed phenotype and the (possibly unknown) phenotype-associated SNP(s) genotype. For example, in a "dominant" mode of inheritance, the phenotype would be expressed whenever the causal allele was present in the associated SNP, regardless of homo/heterozygosity. In contrast, a recessive mode would result in an expressed phenotype only when the SNP was homozygous in the causal allele, and an additive mode would result in a phenotype expression proportional to the number of causal alleles present. All this can be quantified using relative weights as in Table 27.12. This table can be augmented with your own custom modes (*Tools* → *Options* → *Miscellaneous* → *Mode of Inheritance*)

Mode	AA	AB	BB
Additive	0	1	2
Recessive	0	0	1
Dominant	0	1	1

Table 27.12: Mode of Inheritance

27.6.2 Statistical Analysis

The tests in this section aim to answer the central question of association analysis: *which SNPs are associated with expression of the phenotype being studied?* They are all statistical hypothesis tests. The calculation details are in the tables below. Some tests have additional options which may be selected:

- Mode of Inheritance: Can be selected only if the test's calculation provides for a flexible assignment of numbers to genotypes
- EIGENSTRAT corrected data: Possible only if the test makes no assumption that genotypes/phenotypes must be integer-valued
- Apply **Genomic Control**: Possible only if the test is a χ^2 test
- Permutative p-value computation: Possible for all those that are not exact tests
- Multiple testing correction: Always possible, because every test generates a p-value

Brief descriptions of the tests are as follows.

Pearson's χ^2 test:

Tests whether there is a significant correlation between the allele counts of an SNP and case-control status. It can only be applied to case-control experiments, and cannot be applied to EIGENSTRAT-corrected data. For some SNPs, this automatically gets replaced by Fisher's exact test (refer to note 27.6.2, and section 27.6.2). It is possible to run genomic control on the χ^2 test.

Note: If any of the counts, $E_{ij} \leq 10$ (**What is E_{ij} ?**) for an SNP, **GeneSpring GX** calculates the p-value using the Fisher's exact test.

Fisher’s exact test:

A more accurate version of Pearson’s χ^2 test retaining all the characteristics of that test. Genomic control cannot be run on this. Neither can permutation testing.

Cochran-Armitage test:

Tests whether there is a significant correlation between the genotype counts of an SNP and case-control status. It can only be applied to case-control experiments, and cannot be applied to EIGENSTRAT-corrected data. It is possible to run genomic control on this.

 χ^2 correlation test:

Tests whether there is a significant correlation between the genotypes and phenotypes across samples at a given SNP. It can be applied to any data, including EIGENSTRAT-corrected data. It is possible to run genomic control on this.

Note: When an SNP has the same genotype for all the samples all the statistical analysis tests would set the p-value to 1; such an SNP cannot be associated with any trait

- **Entity List and Interpretation (Step 1)**

Select an Entity List and Interpretation from the Entity List and Interpretation dialog, and click Next. This opens the Select a Test dialog.

- **Select a Test (Step 2)**

Select a test, the cases, and controls from the respective drop-down lists, and click Next. This opens the Select Inputs dialog.

- **Select Inputs (Step 3)**

- Select a Mode of Inheritance (default option: Additive; refer to Table 27.12) from the drop-down list.

Note: You cannot select the Mode of Inheritance for χ^2 and Fisher’s Exact tests.

- You can select the Use Corrected Data check box to run the χ^2 Correlation test with the data corrected for stratification.

Note: This option is only available for χ^2 Correlation test.

- Select a method for p-value Computation (default option: Asymptotic).

Note: You can enter the Number of Permutations (default value: 100) for the Permutative option.

- You can select a Multiple Testing Correction method (default: No Correction; refer to section 27.6.2), and click Next. This opens a list of significant SNPs and the respective p-values [7], [53].

- You can select the Genomic Control (refer to section 27.6.2) check box for the Pearson’s χ^2 , χ^2 Correlation, and Cochran-Armitage tests [44].

Note: This option is not available for Fisher’s exact test.

- **Significant SNPs (Step 4)** Significant SNPs dialog launches a list of Significant SNPs with their respective p-values.

Note: Click on the Change Cutoff button to change the p-value cutoff from the slider or text box.

Genomic Control

Genomic Control [44] is a standard method to compensate for population stratification (refer to section 27.4.3). It is a modification to the calculation of a χ^2 test. When there is stratification present in the samples, the test statistics (χ^2 tests) for all the SNPs get inflated by some constant factor (> 1); the less the stratification, the closer the factor is to 1. Genomic control attempts to measure this factor and downscale all calculated χ^2 statistics by this factor to mimic what the results would have been without stratification.

GeneSpring GX offers two approaches to deal with stratification: **EIGENSTRAT correction** and Genomic Control. EIGENSTRAT is the more rigorous and preferable method in most GWAS situations. However, there are two situations in which Genomic Control may be preferred:

- If the amount of data is too small to confidently run EIGENSTRAT (refer to section 27.4.3).
- If the Pearson's χ^2 or Cochran-Armitage test is desired (since these do not run on EIGENSTRAT corrected data, but do accept Genomic Control).

Permutative p-value Computation

All the tests (except for Fisher's exact test) offered here make standard statistical assumptions that approximate the exact answer better when run with more data. However, the results are only asymptotically accurate. The accuracy is comprised in the case of small- to moderate-sized datasets. Permutative testing solves this; it does not hinge on statistical assumptions and essentially estimates the probabilities involved through repeated simulation. The simulation can sacrifice computing power for accuracy.

Number of Permutations: It determines how accurate the answer is. The default (100) is generally acceptable. The higher the setting, the more accurate the answer, but longer is the time taken on a machine.

Multiple Testing Correction

Multiple Testing Correction (MTC) is a statistical problem in GWAS that is addressed using one of a family of methods that **GeneSpring GX** provides.

- More stringent Multiple Testing Correction (MTC) methods declare less SNPs significant, but have a lower risk of false positives. **GeneSpring GX** offers MTC methods with varying levels of stringency, which depends on the data to some extent. A popular ordering of these methods from most to least conservative is as hereunder:
 1. Benjamini-Yekutieli (BY)
 2. Bonferroni
 3. Benjamini-Hochberg (BH)
 4. Storey's q-value
- Less stringent methods are recommended for first-stage analysis to avoid false negatives. BY, BH, and Storey's (Storey's in particular) perform better when there are more truly significant SNPs.

Quantitative and Categorical Trait Types

GeneSpring GX implicitly classifies a trait as either categorical or quantitative. Quantitative traits are those for which the interpretation parameter value is numeric for all samples. Categorical traits are all others.

Categorical traits require a Binary/Nominal/Ordinal trait type selection to be done before running some tests. Also, they result logistic regression being run when the Regression wizard is run. Linear regression is run instead for quantitative traits.

Pearson's χ^2 Test

GeneSpring GX allows you to perform χ^2 test based on disease status of each sample, and genotype information for all SNPs, across samples.

Calculation

- Suppose the allele A and B counts for Cases and Controls are N_{AC} , N_{BC} , N_{ACont} , and N_{BCont} respectively.
- The expected occurrences of missing and non-missing genotype calls are given by the relations: $E_{AC} = (N_A \times N_C)/N$, $E_{BC} = (N_B \times N_C)/N$, $E_{ACont} = (N_A \times N_{Cont})/N$, and $E_{BCont} = (N_B \times N_{Cont})/N$.
- $\chi_{AC}^2 = (N_{AC} - E_{AC})^2/E_{AC}$. Similarly, χ_{ACont}^2 , χ_{BC}^2 , and χ_{BCont}^2 are calculated.
- $\chi^2 = \chi_{AC}^2 + \chi_{ACont}^2 + \chi_{BC}^2 = \chi_{BCont}^2$

Input	<ul style="list-style-type: none"> • Genotype information for all SNPs, across samples • Disease status for each sample • Interpretation with two conditions • Genomic Control (optional; refer to section 27.6.2) • p-value Computation method (default option: Asymptotic) • Multiple Testing Correction (optional; refer to section 27.6.2)
Process (for each SNP)	<ul style="list-style-type: none"> • Calculates the allele counts for each SNP, combined across samples. • Calculates the χ^2 statistic for each SNP. • Finds the p-values, given the χ^2 statistic is calculated with one degree of freedom. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values.

Table 27.13: Summary of Steps: Pearson’s χ^2 Test

Note: If any of the counts, $E_{ij} \leq 10$ for an SNP, **GeneSpring GX** calculates the p-value using the Fisher’s exact test.

	Allele A	Allele B	Total
Cases	N_{AC}	N_{BC}	N_C
Controls	N_{ACont}	N_{BCont}	N_{Cont}
Total	N_A	N_B	N

Table 27.14: Contingency Table for Pearson’s χ^2 Test

Fisher’s Exact Test

GeneSpring GX allows you to perform Fisher’s exact test based on disease status of each sample, and genotype information for all SNPs, across samples. Fisher’s exact test is typically not used for a relatively large number of samples. There are two reasons for this:

- Fisher’s exact test can take a significant amount of computing power if the number of samples is very

high. The definition of "very high" depends on your machine specification.

- Pearson's χ^2 test becomes asymptotically identical to Fisher's exact test for a large number of samples

Input	<ul style="list-style-type: none"> • Genotype information for all SNPs, across samples • Disease status for each sample • Interpretation with two conditions • Multiple Testing Correction (optional; refer to section 27.6.2)
Process (for each SNP)	<ul style="list-style-type: none"> • Calculates the allele counts for each SNP, combined across samples. • Calculates the allele count for each SNP, combined across samples. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values.

Table 27.15: Summary of Steps: Fisher's Exact Test

Cochran-Armitage Test

GeneSpring GX allows you to perform the Cochran-Armitage test for trend. This tests if there is a correlation between genotype counts for a given SNP and case-control status. More specifically, it tests for a linear trend between the genotype and phenotype.

χ^2 Correlation Test

GeneSpring GX allows you to perform χ^2 correlation test. This statistically tests the value of Pearson's correlation coefficient (r) between genotypes and phenotypes.

Input	<ul style="list-style-type: none"> • Genotype information for all SNPs, across samples • Disease status for each sample • Interpretation with two conditions • Mode of Inheritance (default option: Additive; refer to section 27.6.1 section) • Genomic Control (optional; refer to section 27.6.2) • Multiple Testing Correction (optional; refer to section 27.6.2)
Process (for each SNP)	<ul style="list-style-type: none"> • Calculates the Genotype call counts for each SNP, combined across samples. • Calculates the test statistic for each SNP using χ^2 distribution (one degree of freedom). • Finds the p-value for each SNP. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values.

Table 27.16: Summary of Steps: Cochran-Armitage Test

	Genotype AA	Genotype AB	Genotype BB	Total
Cases	r_0	r_1	r_2	r_+
Controls	s_0	s_1	s_2	s_+
Total	n_0	n_1	n_2	N

Table 27.17: Contingency Table for Cochran-Armitage Test

Mode	Genotype AA (d_0)	Genotype AB (d_1)	Genotype BB (d_2)
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1

Table 27.18: Weights (d_i) for Cochran-Armitage Test

Input	<ul style="list-style-type: none"> • Entity List of SNPs • Trait type: Quantitative or Categorical (refer to section 27.6.2). Note: Categorical trait could be in Nominal or Ordinal scale. • Mode of Inheritance (default option: Additive; refer to section 27.6.1) • Genomic Control (optional; refer to section 27.6.2) • Multiple Testing Correction (optional; refer to section 27.6.2) • EIGENSTRAT corrected data (optional)
Process (for each SNP)	<ul style="list-style-type: none"> • Performs correlation test between the SNP data and the traits for each SNP, across samples. • Generates a χ^2 statistic. • Finds the p-value for each SNP. • Filters out the SNP if the p-value is above the cutoff.
Output	List of SNPs to be kept with respective p-values.

Table 27.19: Summary of Steps: χ^2 Correlation Test

Calculation:

Let x_i and y_i be the genotype and trait for the i^{th} sample.

Test statistic (uncorrected data) =

$$(n-1) \frac{\left(\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{j=1}^n y_j \right)}{n} \right)^2}{\left(\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right) \left(\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} \right)}$$

Let k be the number of PCs along which the data has been corrected.

Test statistic (corrected data) =

$$(n-k-1) \frac{\left(\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{j=1}^n y_j \right)}{n} \right)^2}{\left(\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right) \left(\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n} \right)}$$

Nominal and Ordinal Scales

Nominal Scale:

At the nominal scale one uses labels; for example, eye color is a trait with multiple labels: black, brown, hazel, green, blue, etc. Nominal measures offer names or labels for certain characteristics.

Ordinal Scale:

In this scale type, the numbers assigned to objects or events represent the rank order (1st, 2nd, 3rd, etc.) of the entities assessed. An example of ordinal measurement would be any disease state classified as none, mild, moderate, or severe. Note that in the nominal scale there is no order implied. A "high-medium-low" attribute would be best assigned as ordinal because reordering the attribute values (e.g., to "high-low-medium") makes a difference.

27.6.3 SNP Tagging

GeneSpring GX allows you to identify tag SNPs from an input list of SNPs. These tag SNPs, which often only comprise a small fraction of the original SNP list, can be used on behalf of the original set of SNPs in other analyses without losing much accuracy.

Any non-tag SNP in the list would be well correlated with a tag SNP. In other words, its genotype value could be predicted with good accuracy by the tag SNP's genotype value. So analyzing all the SNPs, and

investigating which of them is associated with the phenotype, is equivalent to analyzing just the tag SNPs. This can take much less time and computing resources, as well as minimize the multiple testing problem.

Carlson's greedy algorithm [13] is being run to identify the tag SNPs. The main parameters associated with this algorithm are described hereunder:

MAF Cutoff:

The tagging algorithm can give unpredictable results if any allele frequency is too low. To prevent this, the algorithm has a self-contained MAF filter exactly like the one mentioned in the "Filters" section (27.5.4).

LD cutoff:

r^2 is a commonly used measure of correlation between two SNPs; it is just the square of Pearson's correlation coefficient (r) between the two sets of genotype calls across samples. As such, it can range from 0 to 1, with better correlation being a higher r^2 . Our tagging method can guarantee that any non-tag SNP in the input list will have an r^2 with some tag SNP of at least this LD cutoff. So every non-tag SNP will be guaranteed to be represented in the tag SNP list, to an extent measured by this cutoff. A higher cutoff results in more tag SNPs (higher cost of analysis), which represent the data better, i.e., the information loss from excluding the non-tag SNPs will be lower. A lower cutoff causes the opposite.

Distance Threshold:

Decides how far away a non-tag SNP can be from the tag SNP that represents it. A higher value here can improve accuracy marginally but will increase the time taken, sometimes drastically. A lower value does the opposite.

Workgroup Users: You can run SNP tagging remotely using the workgroup server; refer to section 30.3.2 for details.

Summary of Steps

This section provides a summary of steps involved in SNP Tagging.

Input

- Entity List of SNPs
- MAF cutoff (default value= 0.05)
- LD (r^2 cutoff (default value= 0.8)
- Distance Threshold (default value= 50)

Process (for each SNP)

- Calculates the MAF for each SNP across all samples.
- Filters out SNPs with MAF below the cutoff.
- Executes Carlsons algorithm on the remaining SNPs [13].

Output

The tag SNPs, a subset of all input SNPs.

Recommendation: You can Tag SNPs after any of the QC, Filter, or Analysis steps except for EIGENSTRAT Correction on Samples.

27.6.4 SNP Regression

GeneSpring GX allows you to perform linear and logistic regression of sample genotypes against phenotypes. You can regress each SNP independently or multiple SNPs simultaneously (none of the individual SNPs in a gene might meet the significance threshold, for example, but the block might when tested as one block).

Summary of Steps

This section provides a summary of steps involved in SNP Regression.

Input

- Entity List of SNPs
- Trait type: Quantitative [6] or Categorical (Binary [20], Nominal, or Ordinal [41]; refer to sections [Quantitative and Categorical Trait Types](#) and [Nominal and Ordinal Scales](#))
- Mode of Inheritance (default option: Additive; refer to section 27.6.1)
- Option for Single or Multiple SNPs
- Multiple Testing Correction (optional; refer to section 27.6.2)
- EIGENSTRAT corrected data (optional)
- p-value Computation method (default option: Asymptotic)

Note: **GeneSpring GX** does not support Permutative method for p-value Computation for Categorical traits.

Process (for each SNP)

Quantitative:

runs multiple linear regression.

Categorical:

runs multiple logistic regression.

Note: "Reference Phenotype" refers to the de facto **control**-like phenotype value (for a nominal trait), with respect to which all tests are run. For example, if a disease is being studied and there are 6 phenotype values for a nominal trait, 5 will be considered as different phenotypic expressions of disease-affected status, and the other one will be considered as the phenotypic expression of disease-unaffected status.

Output

List of SNPs with χ^2 and T-Statistics p-values. If the trait is quantitative, then F-statistic is used instead of χ^2 .

Note: the p-value is an indicator of the strength of correlation between the disease status and the SNP or multiple SNPs.

Summary of Multiple Linear and Logistic Regression

This section provides a summary of steps involved in Multiple Linear and Logistic Regression.

Input

Genotype information for all SNPs, across samples

Process (for each SNP)

Performs MLR, using each sample as a data point

- Independent variables: SNP genotypes
- Dependent variable: trait being studied

Output

- p-value for each SNP.
- Regression coefficients and associated likelihood statistics for each SNP.

27.6.5 Haplotype Trend Regression

GeneSpring GX allows you to identify if a haplotype [54] several SNPs long is associated with the phenotype, even if none of the individual constituent SNPs are. An algorithm called Haplotype Trend Regression is used to do this. The implementation ranges across the entire input SNP list, testing haplotype blocks (groups of SNPs) of a fixed length that you can specify. The feature allows the "window" containing the SNPs (being tested) to slide along each chromosome, testing individual haplotypes within each block. The windows can only slide along one chromosome; switching to another chromosome starts the process again.

This step first infers all relevant haplotype frequencies in the "sliding windows" to be tested. Then, for each window, it calculates a t-statistic p-value for each possible haplotype that can occur in the window (using the calculated haplotype frequencies), as well as one F-statistic p-value for the entire window (Refer to [Haplotypes view](#) section).

If the window size was 3 and there were 5 input SNPs numbered S1,S2,...,S5 respectively, the algorithm would test all possible haplotypes of SNPs S1,S2,S3; then test all possible haplotypes of SNPs S2,S3,S4; and lastly do the same for SNPs S3,S4,S5.

Summary of Steps

This section provides a summary of steps involved in haplotyping trend regression.

Input

- Entity List of SNPs
- Trait type: Quantitative [6] or Categorical (Binary [20], Nominal, or Ordinal [41]; refer to sections on [Quantitative and Categorical Trait Types](#) and [Nominal and Ordinal Scales](#))
- Number of Loci (default value: 3)

Note: it is the size of the sliding window, and the length of all haplotypes tested.

Process (for each SNP)

- Identifies the type of trait

Quantitative:

runs multiple linear regression.

Categorical:

runs multiple logistic regression.

Note: "Reference Phenotype" refers to the de facto control-like phenotype value (for a nominal trait), with respect to which all tests are run. For example, if a disease is being studied and there are 6 phenotype values for a nominal trait, 5 will be considered as different phenotypic expressions of disease-affected status, and the other one will be considered as the phenotypic expression of disease-unaffected status.

- Generates haplotypes and their frequencies.
- Filters out haplotypes with frequencies less than 0.01, and displays the haplotypes and their frequencies.
- Runs a regression test on the remaining haplotypes, and generates the p-value for each Haplo block.

Output

- Provides a Haplo block list. The first SNP of the block will be stored in the list.
- Provides F-statistic p-value (for Quantitative traits) for each Haplo block. If the trait is categorical, then χ^2 is used instead of F-statistic .
- Lists all the haplotypes with respective t-statistic p-values.

Note: HTR entity List is a special list that cannot be used for downstream analysis. Contact GeneSpring support for a script to convert an HTR list to a 'regular' entity list for downstream analysis.

Note: you can configure the Haplotyping Parameters from the tools menu.
(*Tools* → *Options* → *Miscellaneous* → *Haplotyping Parameters*)

Haplotypes view

In **GeneSpring GX** you can launch the Haplotypes view from the Haplotype Entity List Inspector.

The view launches a list with the following columns:

Probe set Id or Name:

Provides the Name (Illumina) or Probe set id (Affymetrix) of the first SNP in the Haplo block.

F-Statistics p-value:

Provides F-statistic p-value for each Haplo block.

Haplotypes:

Lists all the haplotypes for each Haplo block.

T Statistics p-value:

Provides t-statistic p-values for each haplotype.

Haplotypes Context Menu

You can perform common tabular operations using the context (right-click) menu options, which are listed hereunder:

Select All Rows:

Allows you to select all the rows from the list, and then export the view as an image or html file.

Invert Row Selection:

Allows you to invert the row selection, and then use the "Limit to Row Selection" option to launch the selected rows in the view.

Clear Row Selection:

Allows you to clear the existing row selection.

Limit to Row Selection:

Allows you to launch the list with only the selected rows.

Copy View:

Allows you to copy the view to the clipboard.

Print:

Allows you to launch the view in the web browser, which

Export As:

Allows you to Export the view as an Image or HTML file:

- **Image:** Exports the view as an image in .tiff, .bmp, .jpg, .jpeg, .png, or .gif format.
- **HTML:** Exports the view as an HTML file.

Properties:

Allows you to add a Title and Description for the view.

Notes:

- **Quantitative Trait: GeneSpring GX** runs a Multiple Linear Regression.
- **Categorical Trait: GeneSpring GX** runs a Multiple Logistic Regression.
- **EIGENSTRAT corrected data:** You can select the Corrected Data check box to use the EIGENSTRAT corrected data. You must run EIGENSTRAT correction before using this option.

27.6.6 LD Analysis

GeneSpring GX allows you to perform Linkage Disequilibrium (LD) analysis as a method for genetic mapping. LD is the non-random association of two or more loci.

Note: In **GeneSpring GX** LD Analysis runs separately for each Chromosome.

Summary of Steps

This section provides a summary of steps involved in LD Analysis.

Input

- Entity List of SNPs
- Interpretation

Process (for each SNP)

- Filters out SNPs with MAF less than 0.05.
- Launches a list of passed SNPs with respective MAFs
- Calculates the coefficients (r^2 and $D - prime$) of linkage disequilibrium

Output

Creates a subentity list as a **child** under the input entity list in the experiment navigator. If the MAF output is spread across different Chromosomes, then a "child" subentity list is created for each Chromosome (Figure 27.4).

Notes:

- If the no. of SNPs for any Chromosome in the MAF output is greater than the "Maximum no. of SNPs per Chromosome" then Chromosome is not considered further in LD analysis. You can configure the "Maximum no. of SNPs per Chromosome" value from *Tools* → *Options* → *Data Analysis Algorithms* → *LD*.
- If the no. of entities for all the Chromosomes in the MAF Output exceeds the "Maximum no. of SNPs per Chromosome" select *Tools* → *Options* → *Data Analysis Algorithms* → *LD* to configure the value of "Maximum no. of SNPs per Chromosome".
- If the MAF output is spread across different Chromosomes, then the LD Inspector separately lists each Chromosome.

Launching LD Plot

- Click on an LD Plot node in the experiment navigator to launch the LD Plot.

Note: If the selected entity list has entities from multiple Chromosomes then separate LD Plot nodes are created for each Chromosome.

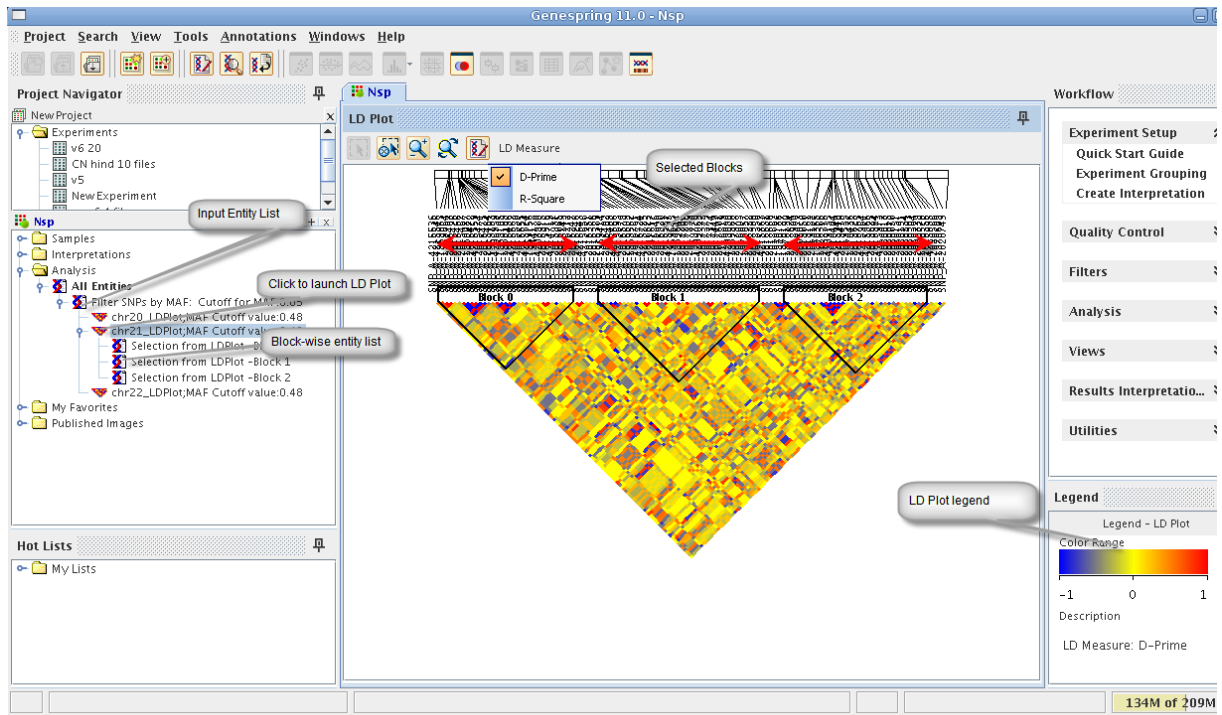


Figure 27.4: LD Plot

- Select an option from the LD Measure menu options: r^2 or $D - prime$ (default option).
- Drag the mouse pointer over the plot to select the blocks of interest (Figure 27.4). Refer to [LD Plot](#) section for more information.

27.7 Views

GeneSpring GX provides an addition tool to view the results of association analysis experiments. You can launch the Genome Browser from the main toolbar or the Workflow link, and drag and drop the results of an association analysis experiment into it.

27.7.1 Genome Browser

You can launch the genotype calls or allele frequencies on the Genome Browser (Figure 27.5). Genotype call to Allele mapping is as illustrated in the table below:

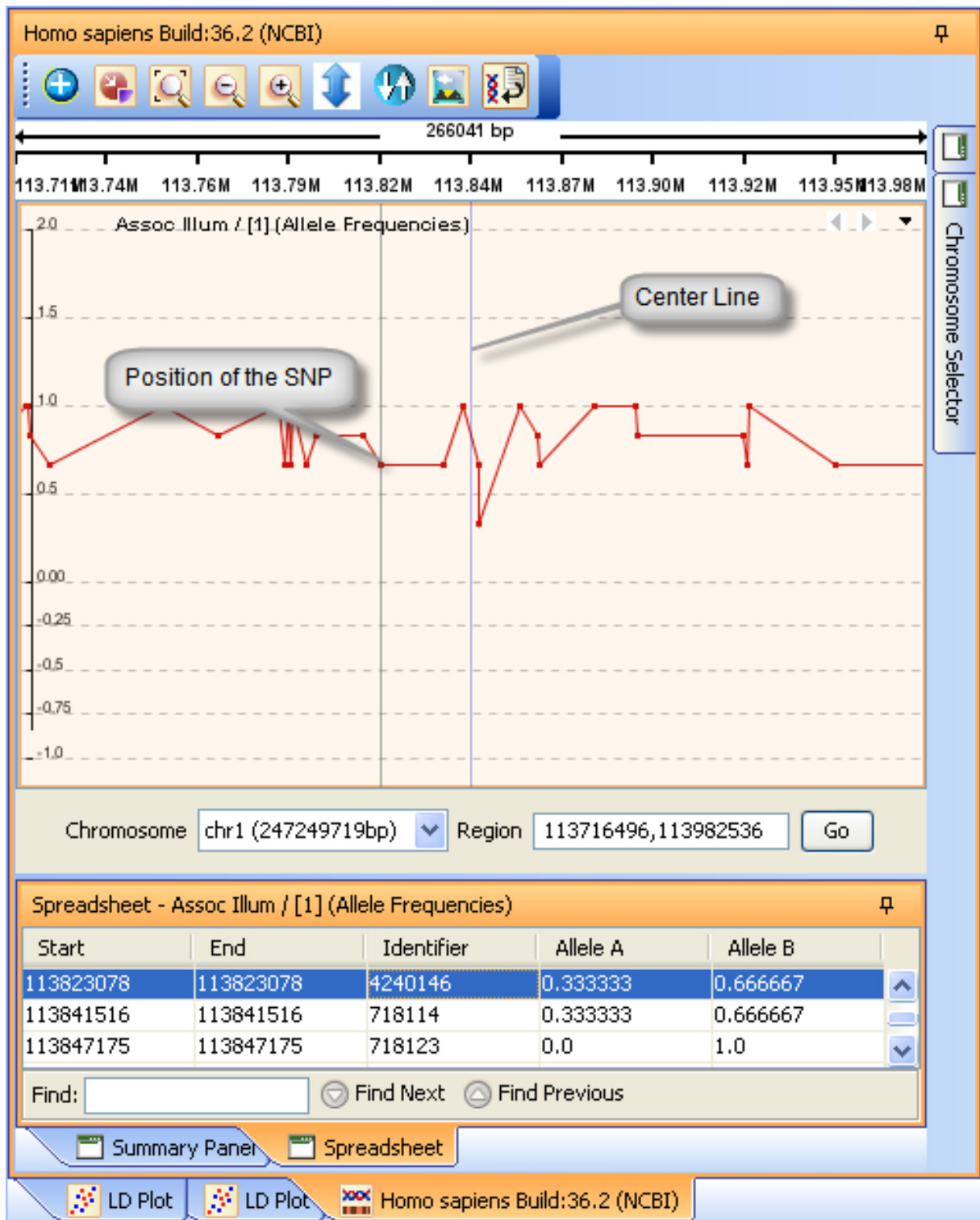


Figure 27.5: Allele Frequencies on Genome Browser

Genotype Call	Allele
-1	No Call
0	AA
1	AB
2	BB

Table 27.20: Mapping for Genotype Calls

Note: You are advised to run the "Calculate Allele Frequency" Utility on an interpretation before launching the Genome Browser

Refer to [Genome Browser](#) chapter for more information.

27.8 Results Interpretations

GeneSpring GX allows you to Identify Overlapping Genes, and run [Gene Ontology](#) and [Pathway](#) Analyses.

27.8.1 Identify Overlapping Genes

GeneSpring GX allows you to identify adjacent genes whose coding regions are partially overlapping.

Input

Entity List of SNPs

Process (for each SNP)

- Identifies a list of genes that are associated with the given SNPs.
- Creates a list of genes that show overlapping.

Output

A list of overlapped genes.

27.9 Utilities

GeneSpring GX provides additional utilities to Import Entity List from a File, Create a Probe List, Calculate Allele Frequency, and Filter on an Entity List.

Import Entity List from File	You can import a file as an Entity List by matching column names. Refer to section Import Entity List from File for details.
Create Probe List	You can create Subset Entity Lists from existing entity list. You can add (and remove) multiple conditions search the existing entity list. Each condition requires Chromosome number, Physical Position, Value, and Metrics. The result is a Probe list matching the conditions.
Calculate Allele Frequency	You can run this utility to calculate Allele frequencies on an interpretation. You can launch the Genome Browser, and then drag and drop the data. This opens the Select Data dialog, which allows you to launch Allele Frequencies along the genome.
Filter on Entity List	You can filter an Entity list using its annotations and list associated values. See section Filter on Entity List for details.

Table 27.21: Utilities in Association Analysis

27.9.1 Using disc cache

Under the menu **Tools** → **Options** → **Copy Number Algorithms** → **APT Execution Options**, 'Use Disc Cache' is an option. It is advisable to use the disc cache in Windows system. Note that it will fail with Linux Operating System with more than 90 samples. When disc cache is not used, the execution process creates batches of probesets and this can slow down the process. In worst situation, the process can quit if there is memory outage. Using disc cache helps to avoid these problems and uses the disc memory while running processes.

Chapter 28

The Genome Browser

28.1 Introduction

Genome Browser in **GeneSpring GX** is a powerful visualization aid to view all kinds of data associated with the genome in one place. It can even be used to visualize genome data available outside the **GeneSpring GX**, provided the chromosome start and end detail is there in the file. With the Genome Browser, user can

- Visualize data from different experiments and technologies in one view; however all the data must be from the same organism.
- Visualize data from entity lists.
- Merge different data tracks and analyse data across technologies/experiments/entities.
- Pan the entire chromosome and drill down to gene level and transcript level for viewing data.
- Import annotation data on the fly for any organism and view it in the Genome Browser along with your own experimental data.
- Export entity lists from the experimental data tracks.
- Publish images and share them across users in workgroup mode; Export plots as images for presentations/publications.
- Drag and drop files containing chromosome information from outside of the tool and view it in the genome browser.
- View the associated data of the selected region in the chromosome in the spreadsheet form; Export selected region for later analysis.

Any experiment that is chosen for visualization in the Genome Browser should contain technology annotations describing the chromosomal location of the entities (expression/exon/SNP probesets/genes) in the experiment. This includes:

1. The chromosome number
2. Start location and stop location. In some cases where the entities are SNPs there will be only one location.
3. Strand information. This is optional. Strand information is essential in displaying gene structure (if 'gene' is chosen as an annotation track). Strand information is also required to see the upstream region of a gene.

For standard technologies supported by **GeneSpring GX**, all the above information are available within the experiment and the genome browser will pick this up on the fly. For custom technologies, user needs to ensure that the experiments contain the above information.

28.2 Tracks in Genome Browser

Data is displayed as 'Tracks' in Genome Browser. A 'Track' is a plot of the chromosomal location in one axis against the chosen data in the other axis. The data could be annotation data of the chromosomes, raw or normalized signal values from expression experiments, copy number or association experiment related data or could even be empty if no data is chosen. Genome browser can show any number of tracks simultaneously, one stacked above the other.

Genome Browser shows Experimental Data Tracks and Annotation Tracks, based on the data that populates them.

Experimental Data Tracks: Data comes from the experiments and is specific to samples. It could be any kind of data like expression, copy number, association, etc.

Annotation Data Tracks: Data is at the organism level and is independent of sample information. Supports annotation information for the locations on the genome for gene, transcript, CpGIlands, miRNA (if applicable). Standard gene, transcript and CpGIland annotation tracks of humans, mouse and rat would come prepackaged with the application.

28.2.1 Track functionalities

- User can seamlessly add new tracks by using the '**Manage Genome Browser Data**' feature'. See [Manage Genome Browser Data](#) for details.

- Another generic way of creating tracks is to drag and drop files from anywhere in the file system into the genome browser. The file should contain chromosome name, start and end position and all data associated with the file can then be viewed inside the genome browser. See section [Drag and Drop Files from anywhere](#) for details.
- User can choose an experiment or an entity list within an experiment in **GeneSpring GX** by dragging and dropping them into the genome browser. See sections [Drag and Drop Experiments](#) and [Drag and Drop Entity list](#) for details.
- Section [Track properties](#) explains the options to view data in the tracks including the data columns to view, plot types, colouring, rendering, scaling, sampling and smoothing, etc.
- Simple operations like merging, export, import, etc are explained in section [Track operations](#).


28.3 Visualization in Genome Browser

The genome browser has three panels.

Selection Panel: Shows all the chromosomes in the organism when the tab called *Chromosome Selector* is clicked. The tab called *Annotations* allows choosing any annotation like gene, transcript, CpGIIsland or miRNA to be shown in the genome browser.

Summary / Spreadsheet Panel: Summary panel displays the chosen chromosome (or chromosome 1 by default) and allows selecting a region within the chromosome to view in the track panel. The annotation information is shown on the relevant positions along the chromosome in this panel. When a track is selected, the spreadsheet will show the corresponding data in a tabular form.

Track Panel: This panel shows the data from the selected chromosome. There are many ways of visualizing data within the track panel which include:

- Choose a chromosome from the 'Chromosome Selector' in the *Selection panel* or choose the required chromosome from a drop down inside the track panel which lists all chromosomes for that organism. Data corresponding to the chosen chromosome is shown in the track panel.
- Select a region from the chosen chromosome in the summary panel or by specifying the coordinates in the drop down 'Region' in the track panel and click *Go*. The selected region is zoomed in and shown in the track panel.
- Alternately, press *Shift* and drag the mouse over a region to select and zoom in the track panel. This can also be achieved by pressing the Zoom-in icon inside the track panel 'Zoom in'  icon.
- Pan the entire chromosome by just clicking and dragging with the left mouse button, after selecting a region in the summary panel. Or use the left and right arrows given in the top right corner of each track to move to the next region with data on the chromosome.
- Panning is also possible with the middle mouse button; scrolling up enables panning to the left while scrolling down enables panning to the right.

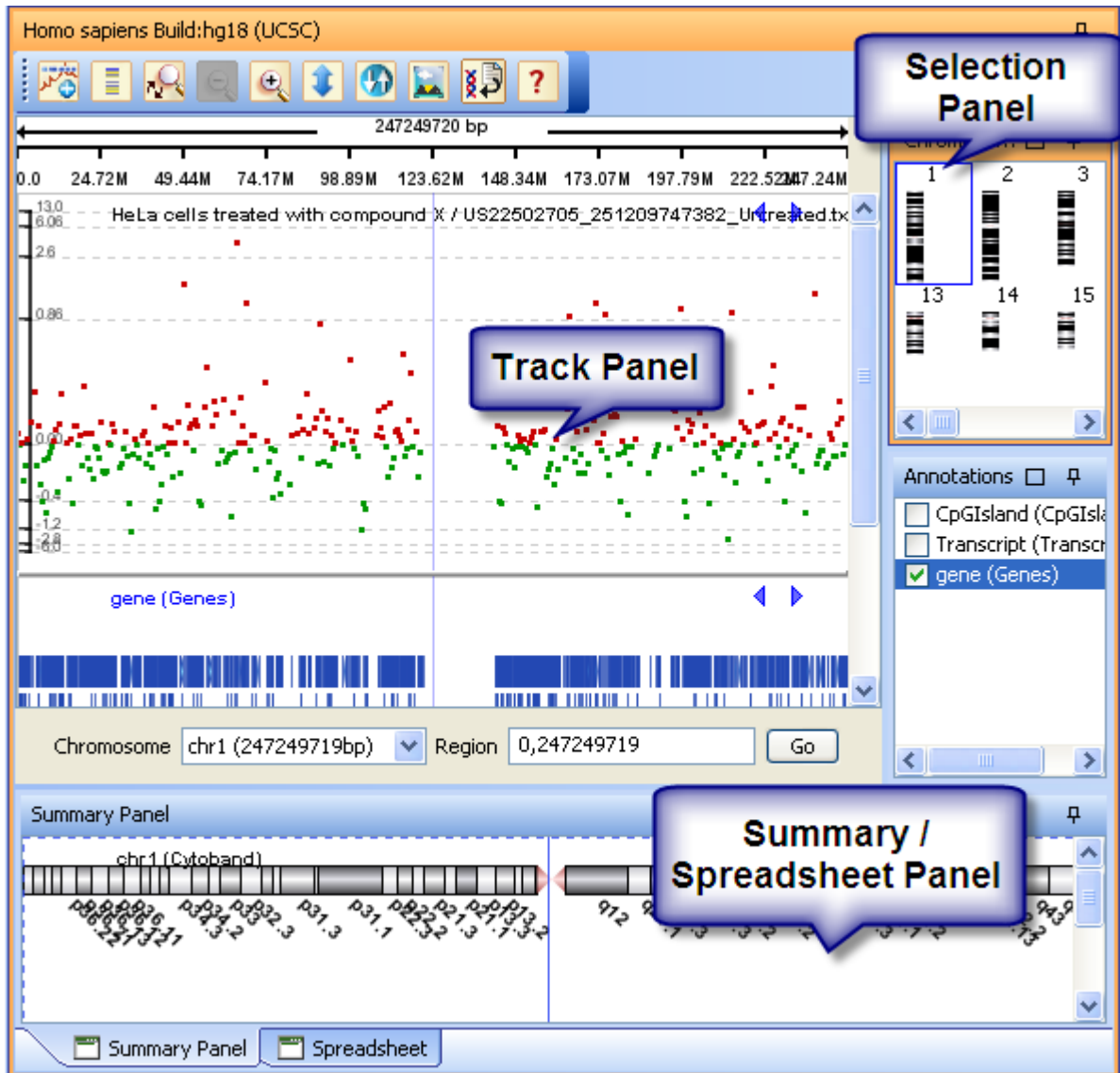


Figure 28.1: Genome Browser showing the panels

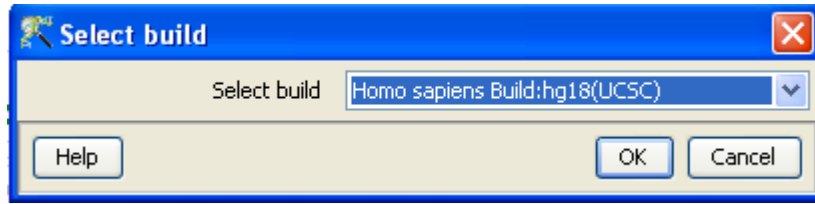



Figure 28.2: Genome Browser - Select Build

- A cross-wire is shown inside the track panel on mouse-over, indicating the position of the chromosome at any point. It can be removed by using the 'Edit Track Properties' option on right click inside the track.
- On right click inside the tracks, there are options to edit the the track properties, to export as entity list or a text file, to remove tracks, to change track size and to split merged tracks. See [Edit Track properties](#) and [Track Operations](#) for more details.

28.4 Working with Genome Browser

Genome browser can be launched from the menu **View** → **Genome Browser** or from the tool bar icon 'Genome Browser'  icon. Note that there should be an active experiment (an opened experiment) to launch the Genome Browser. **GeneSpring GX** comes prepackaged with annotation data for humans (Hg 18 and Hg19 builds), mouse, rat and C elegans obtained from the UCSC site <http://hgdownload.cse.ucsc.edu/downloads.html>; these can be downloaded from the menu **Annotation** → **Update Genome Browser Data** → **From Agilent Server**. It is possible to simultaneously launch multiple builds (which will be present independently). If the active experiment belongs to one of these organism, then a window titled **Select Build** will come up with a drop down to select a build for that organism. User can add new builds and features to these organisms using the [Manage Genome Browser Data](#) functionality explained below.

Choose a build and the genome browser will get launched. On launch, the genome browser shows the annotation information pertaining to chromosome 1 of that particular organism. See [Figure 28.3](#)

Annotation tracks like gene, transcript, CpGIsland can be turned on from the Annotation tab in 'Selection Panel'. User can populate the Genome Browser with other data by any of the following ways:

- [Drag and Drop Experiments](#)
- [Drag and Drop Entity Lists](#)
- [Drag and Drop Files from anywhere](#)

If the organism in the active experiment is not available in the Genome browser and as a download from the menu **Annotations** → **Updata Genome Browser Data**, then a message is shown that 'No Genome

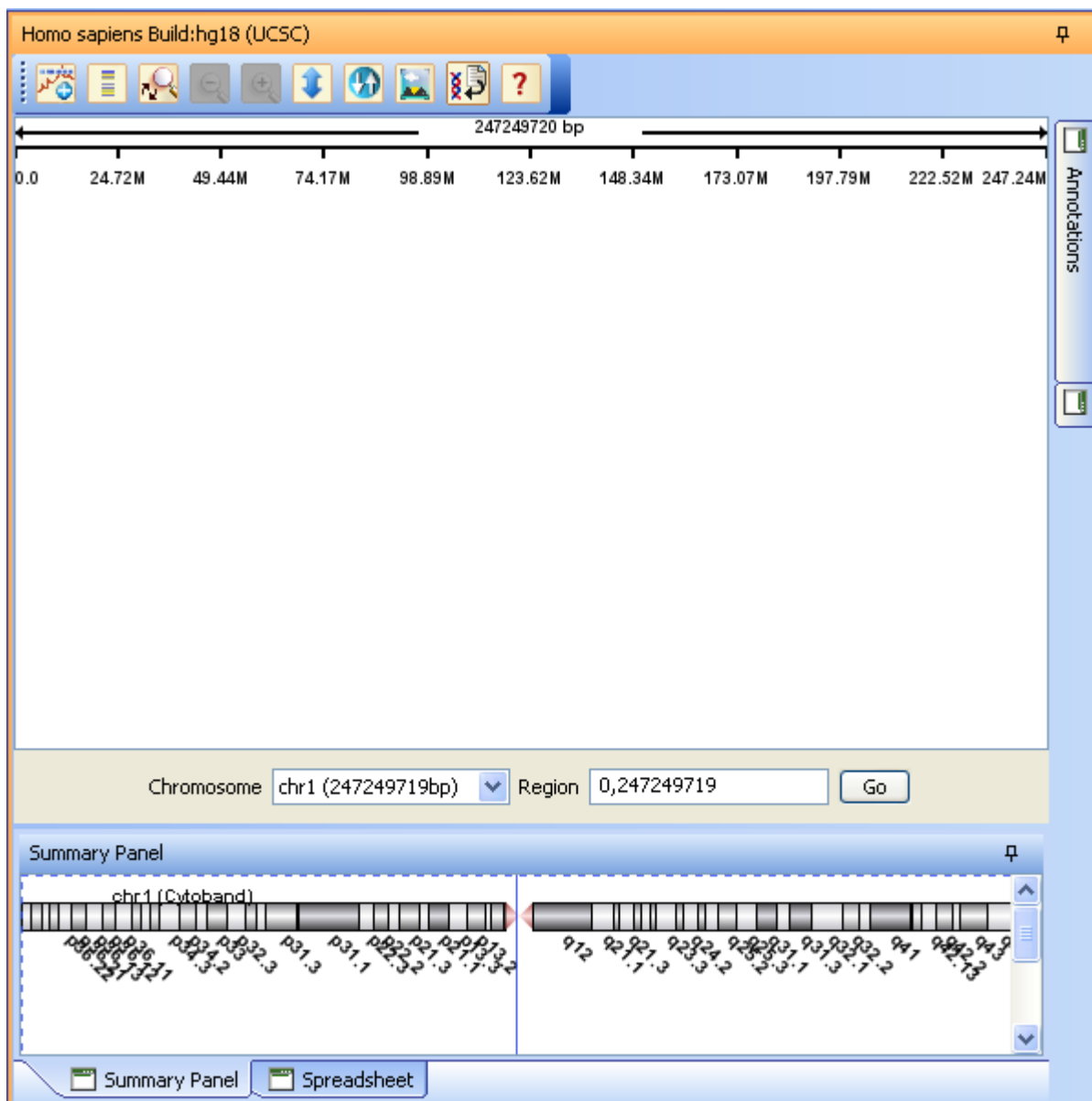



Figure 28.3: Genome Browser - On Launch

builds could be found for the organism'. User can add this organism by using 'Manage Genome Browser Data' from the menu **Annotation** → **Update Genome Browser Data** → **Manage Genome Browser Data** explained in section [Manage Genome Browser Data](#).

28.4.1 Manage Genome Browser Data

This functionality helps to add organisms and builds and also to add new features (annotations) for existing builds. **Manage Genome Browser Data** can be called from the menu **Annotations** → **Update Genome Browser Data** → **Manage Genome Browser Data** or using the tool icon Manage Genome Browser Data  icon, if the genome browser is already launched.

The hierarchy of importing data would be: **Organism** → **Build** → **Features in Chromosomes**

The annotation files can be downloaded from sites like UCSC <http://hgdownload.cse.ucsc.edu/downloads.html>. The files are usually in bed or tab separated formats. While bed files can be directly imported, the other files can be imported using the 'Advanced Import' functionality.

Manage Genome Browser Data brings up a window titled *Import and Manage Tracks*. The browser on the left hand side shows the existing organism, build, chromosomes and features in that hierarchy. Clicking on each of the item will bring up the details in the window on the right hand side.

Additions and deletions can be managed as explained below.

Add an organism: Click on an existing organism on the explorer in the left hand side. On the window that appears on the right hand side, click *New Organism*. Define a Common name, Scientific name and Taxonomy Id for the organism and click *OK*. The new organism will now appear on the explorer in the left hand side. Note that these names are just for identification by the user and need not be technical; the names cannot be edited later though.

Delete an existing organism: Click on the existing organism on the explorer in the left hand side; On the window that appears on the right hand side, click *Delete Organism*.

See Figure [28.5](#) for Organism related operations.

Add Build: From the explorer, choose an existing organism or its build; on the right hand side, click *New Build*. Choose the organism and give a name and source of the build. A date stamp will be automatically generated. New builds can be created using any of the three options explained below; in all of them, the name and length of the chromosome information in files is mandatory.

Figure [28.6](#) shows the interface for adding new build.

Import from tab separated file Any tab separated file containing names of chromosomes along with lengths.

Import from previous builds Any existing build information from the same organism. Choose an existing build from the drop down and there is also an option to add a chromosome and define its length from the interface

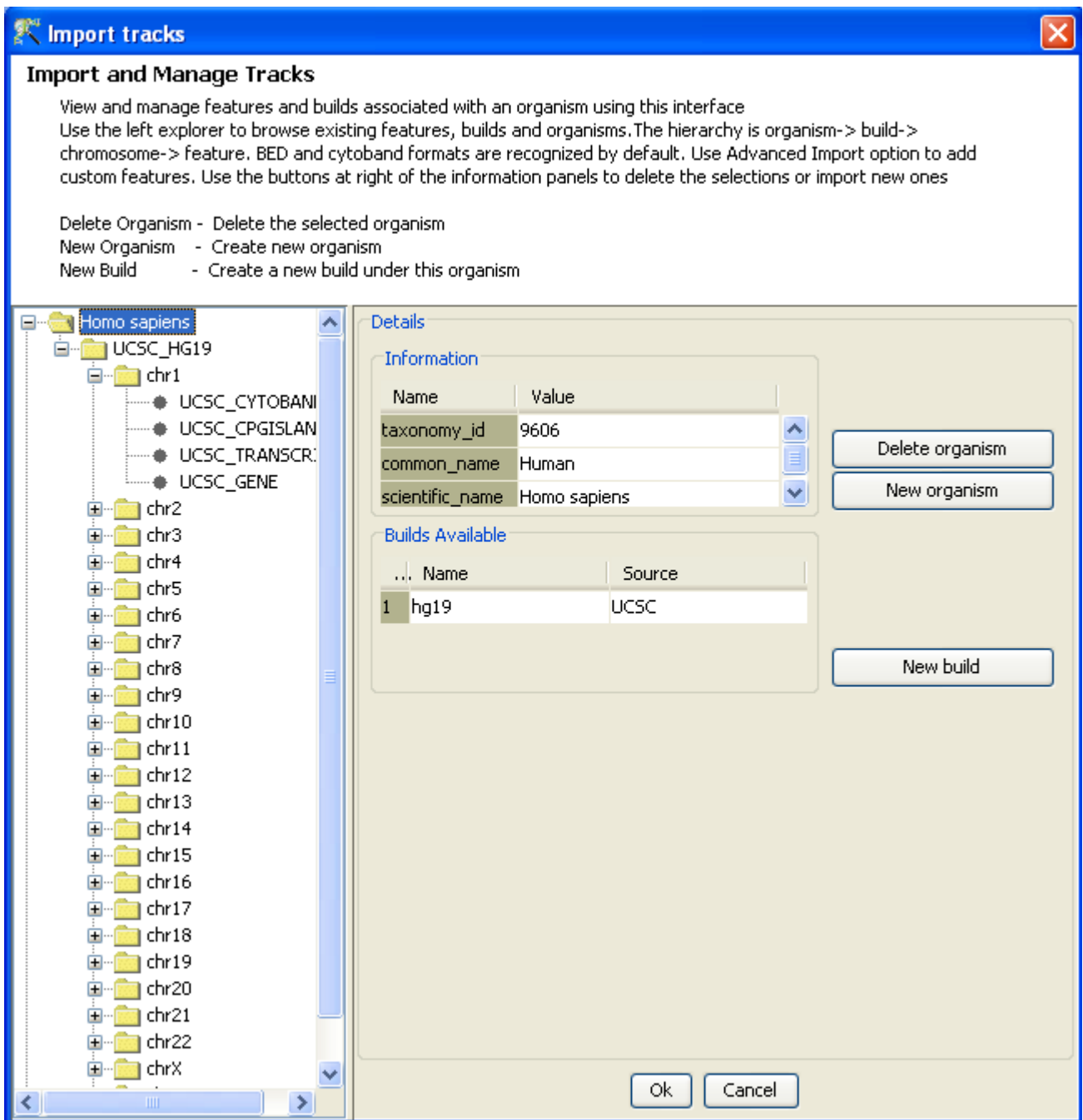


Figure 28.4: Genome Browser - Import and Manage Tracks

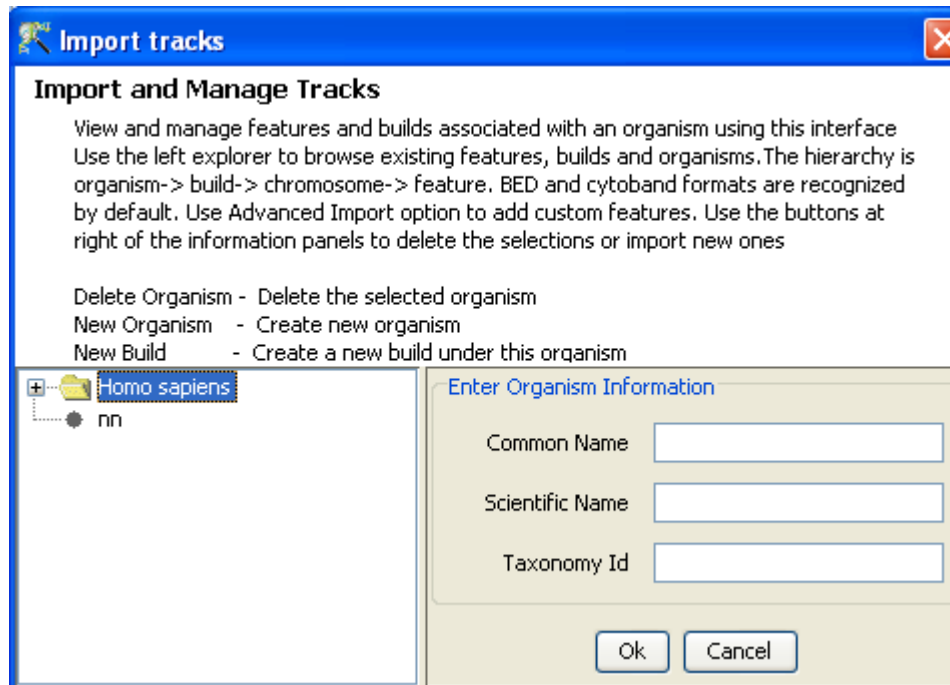


Figure 28.5: Genome Browser - Add/Delete Organism

Import from Cytoband file Use a cytoband file to create a new build.

Delete Build: Click on existing build on the explorer and click *Delete Build* from the right hand side window.

Add Chromosome Feature: To add a new feature, click on a build on the explorer; Click *New Feature* on the right hand side window. A window to define feature information appears. Choose the build, define a name and URL for the feature and input the file containing the features in the file chooser box. If the files are in *.bed formats, it can be directly imported. For files of other formats, use advanced import explained below. A date stamp is automatically created.

Advanced Import for feature: Use this facility to use an existing template or define a new template and import features.

The templates have a predefined format of fields inside the file and optionally, an order for the fields. Templates include:

- Standard annotation files from UCSC including Cytoband, Transcript, miRNA and CpGIIsland are auto-populated with the column types (field names) and the name of the template is the name of the annotation.
- There are generic templates like Segments, SNPs and Genes whose fields are not auto-populated but left to user to fill while importing.
- While working with any of the above templates, user can make changes in the format and save it as a new template; these are User-defined templates and would now appear in the drop down of available templates; User defined templates are also auto-populated with column types.

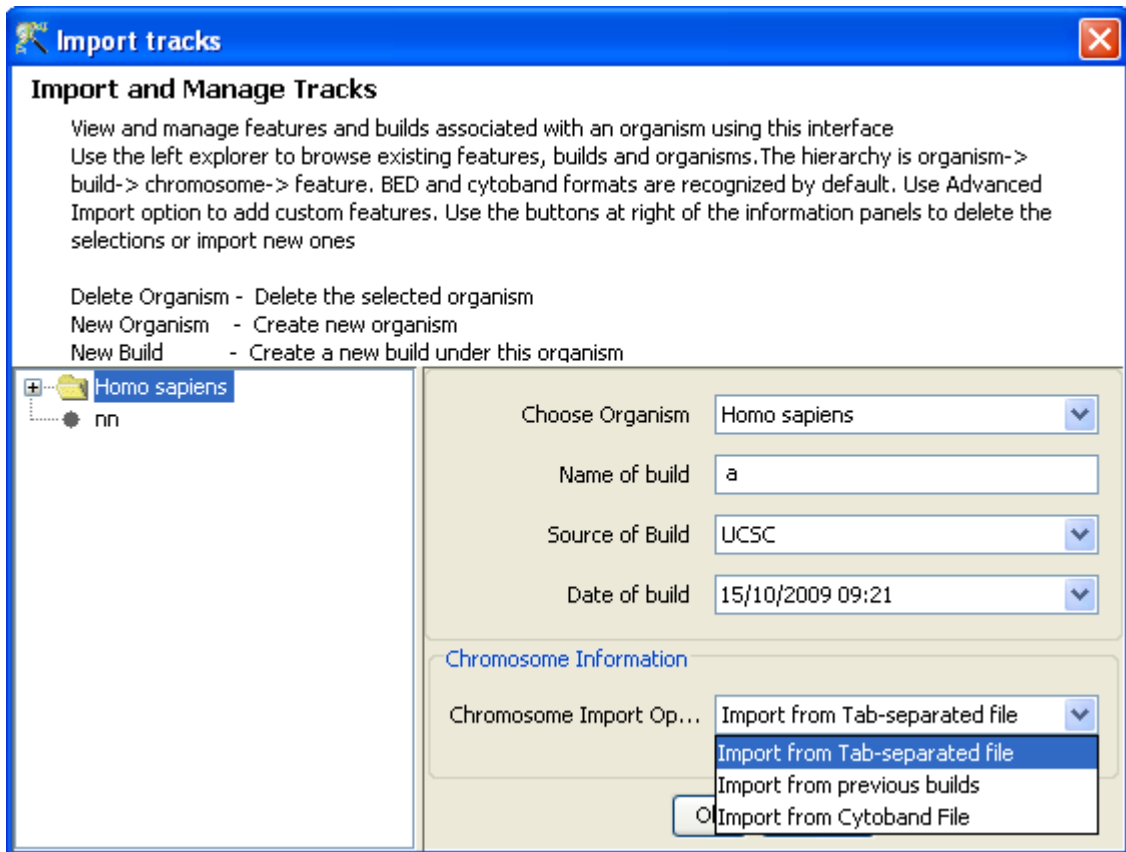


Figure 28.6: Genome Browser - Add New Build

User can import features from various sources as explained below:

Step 1: Feature details: Choose the build of organism and give a name for the feature to be imported. URL field is optional and just for user's reference. Date field is automatically created. Choose the file you want to import and the template. For standard annotation files from sites like UCSC <http://hgdownload.cse.ucsc.edu/goldenPath/>, choose Cytoband, Transcript, CpG-Island or miRNA templates as the case may be. For files exported from **GeneSpring GX** (Export Segments from Copy Number Experiments, for instance), choose a generic template like Segments or SNPs or Genes, whichever is applicable. Similarly, any file containing chromosome name and position can be imported using Segments or SNPs as template.

'Segments' is the default template option as it requires just the Chromosome, start and end details. SNPs needs the chromosome and its position.

Figure 28.7 shows the first step of advanced import using 'Segments' as the template.

Step 2: Format File Define the formatting options like Separator, Text Qualifier, Missing value indicator and Comment Indicator. A preview is shown below along with an option to use the first row as column names.

Figure 28.8 shows the second step of Advanced Import.

Step 3: Choose Columns Define Column types and select the columns to be imported by ticking them. **GeneSpring GX** automatically shows the column types for the standard templates

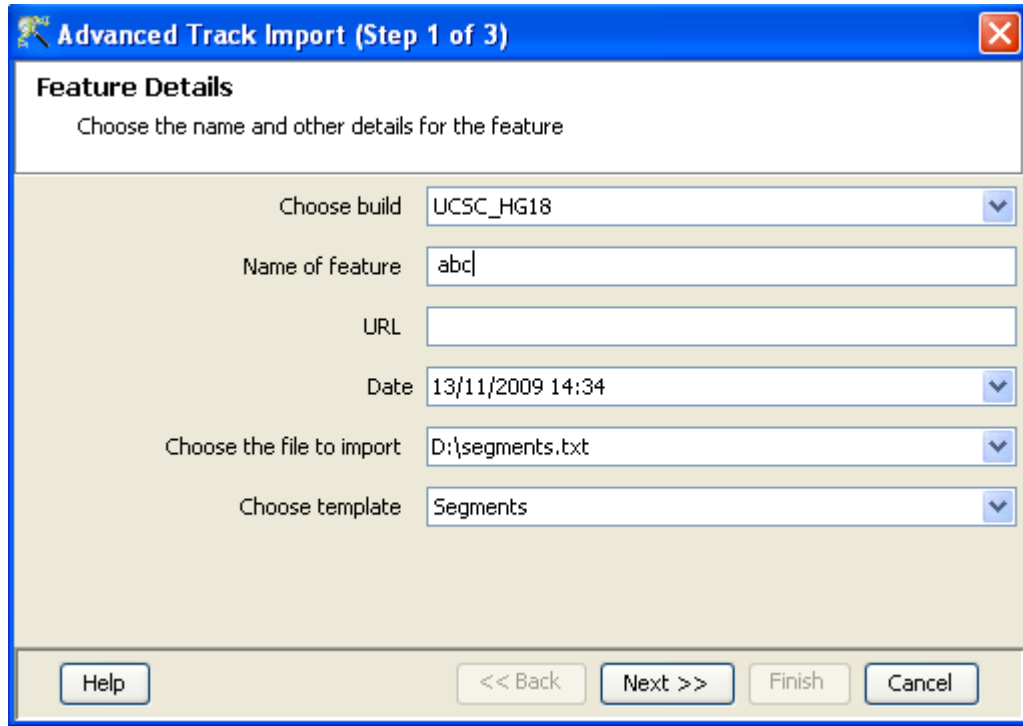


Figure 28.7: Genome Browser - Step 1 of Advanced Import

which include Cytoband, Transcript, CpGIsland and miRNA. User can make changes if required. For generic templates, the columns types are not populated; User needs to fill this from the drop down. A preview of the file is shown below to enable choosing the correct column type. **'Save template as'** option is available to save and use the created templates again.

Figure 28.9 shows the second step of Advanced Import.

Delete Chromosome Feature: Click on an existing chromosome feature and click *Delete Feature* on the right hand side.

28.4.2 Drag and Drop Experiments

Experiments from the navigator panel (which lists all experiments in the project) can be dragged and dropped into the genome browser to populate it. Note that the samples themselves cannot be directly dragged and dropped. A window titled 'Select Data' will appear. On this window, user can choose the following:

1. An *Interpretation*. For expression experiments, if averaged interpretation is used, only the conditions are taken and not individual samples. For copy number experiments, the samples are always taken irrespective of whether an averaged or unaveraged interpretation is chosen. For Association experiments, there is an option to consider genotype calls or allele frequencies; for the former, the samples are taken while for the latter, the conditions are taken.

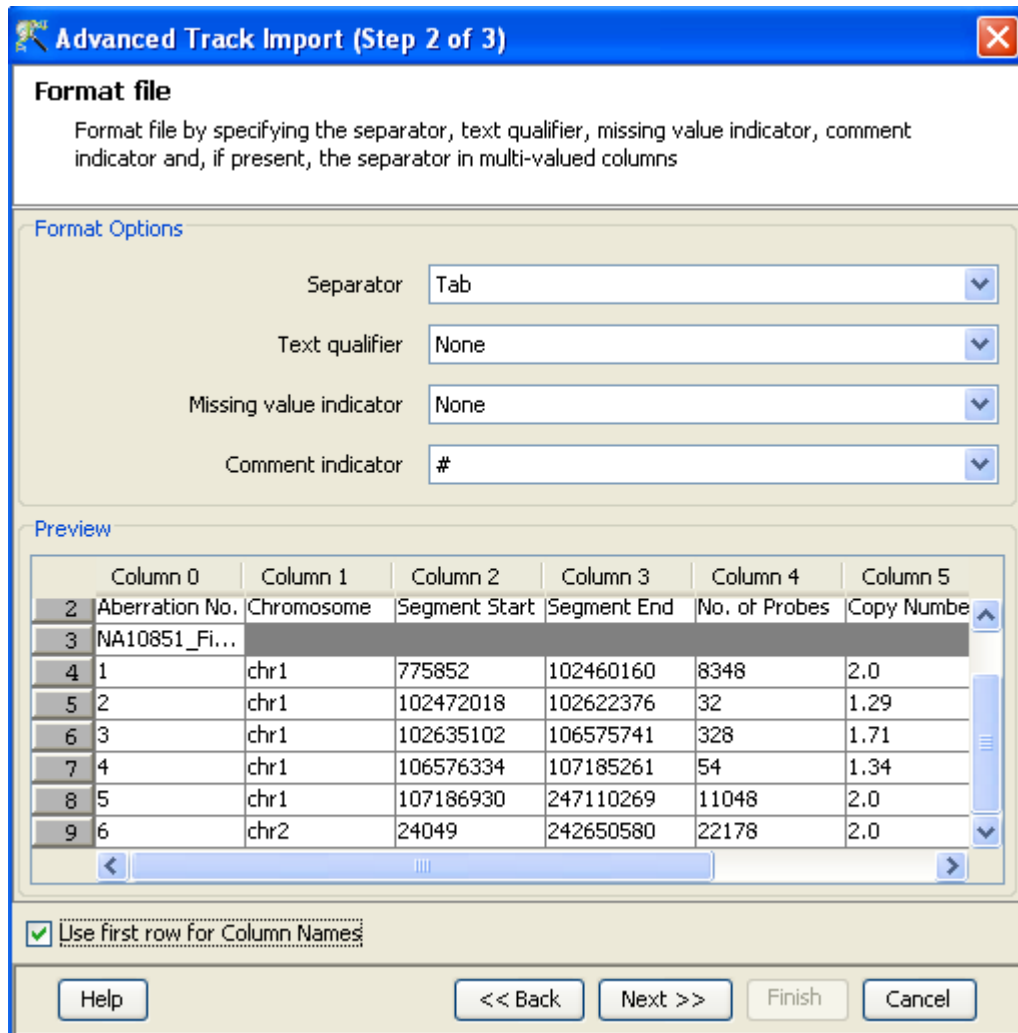


Figure 28.8: Genome Browser - Step 2 of Advanced Import

2. *Samples* Choose samples to be shown in the genome browser. Utilities like *Find*, *Select All* and *Match case* are available to locate samples easily.
3. *Type of data* In experiments other than copy number and association, the raw or normalized (or both) signal values can be viewed in the genome browser. In copy number experiments, user can choose to view the copy numbers, Copy Number confidence, LOH scores, Log Ratios, Allele specific copy number, Parent Specific Copy Number. For association analysis, there is option to view Genotype calls or Allele frequencies. All the data pertaining to that experiment will be listed along with check boxes at the bottom left corner of the window. Check the data that needs to be viewed. For Association experiments, there are tabs to view Genotype Calls and Allele Frequencies.

See Figure 28.10

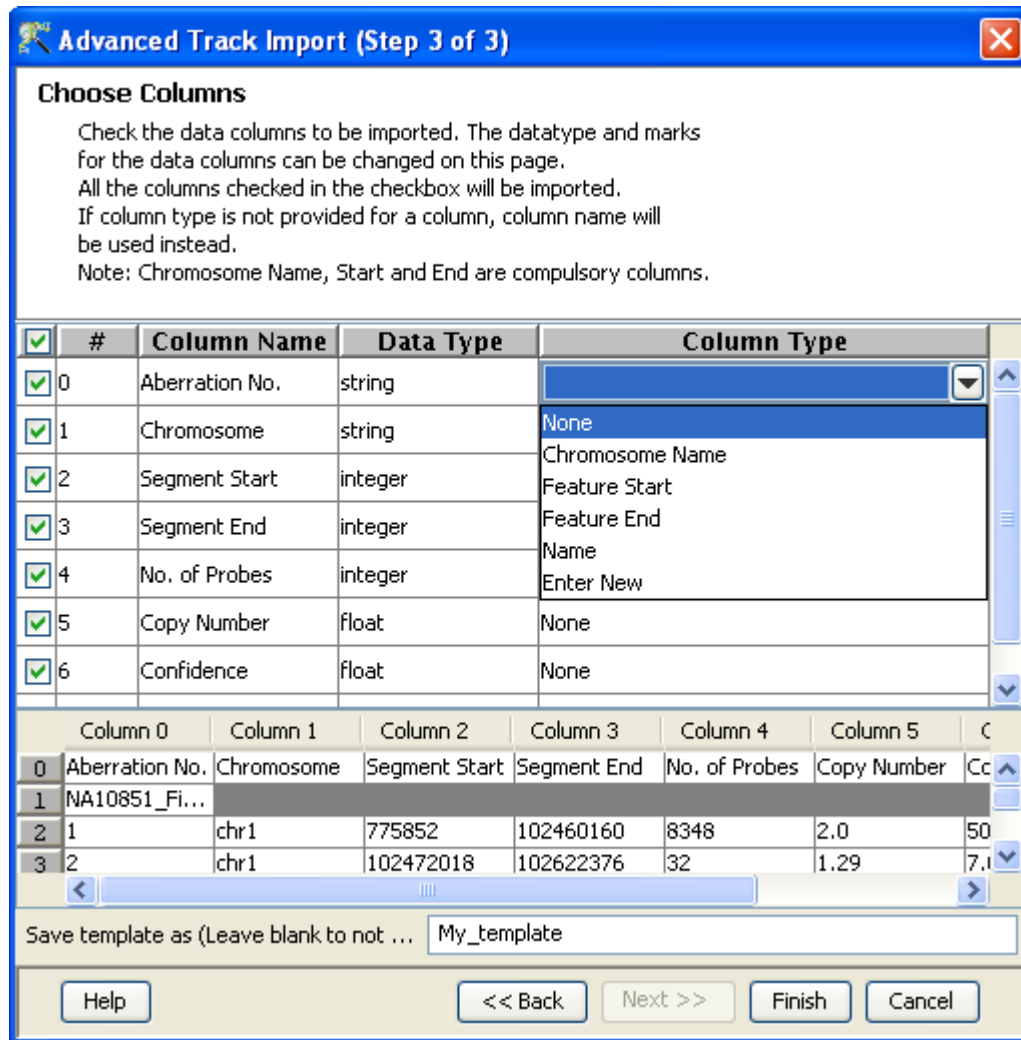


Figure 28.9: Genome Browser - Step 3 of Advanced Import

28.4.3 Drag and Drop Entity Lists

From open experiments, entity list can be dragged and dropped into the genome browser. By default, the track is shown with the location of the entity against the location on the chromosome. Along with this, any other data associated with the entity list can be viewed in the genome browser; choose the required data column through the 'Edit Track Properties' utility.

28.4.4 Drag and Drop Files from anywhere

Files containing minimally chromosome name, start and end location can be dragged from anywhere in the filesystem and dropped into the genome browser. Even if there is a build conflict with an already open experiment, available information from this file can be viewed. Note that these files need not be generated

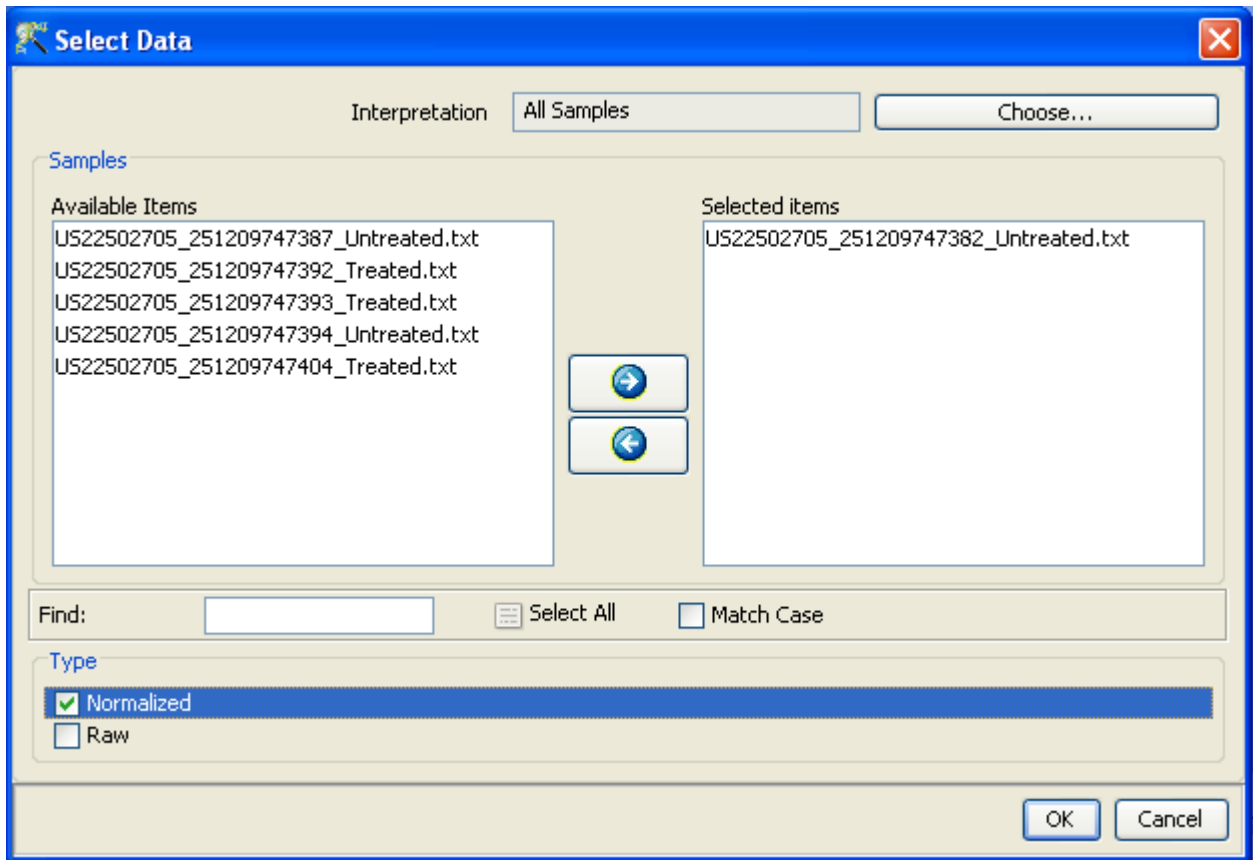


Figure 28.10: Genome Browser - Select Data

from **GeneSpring GX** but the genome browser should have been launched within **GeneSpring GX** (with an active experiment) prior to this dragging and dropping.

A two step wizard comes up to define options while adding the track;


Step 1: Format File This is similar to the advanced import functionality explained in section [Advanced Import](#). Define a name of the feature and choose a template (default 'Segments'). The **'Import track permanently'** option saves this file information as part of the annotation for the particular build of that organism.


User can choose the formatting options here and a preview of the file is shown at the bottom.

Step 2: Choose columns In this step, user can define the column types for import, as in step 3 of [Advanced Import](#). 'Save template as' option is available to save and use the created templates again.



28.4.5 Track Operations

Tool bar icons on top of the Genome Browser view enables carrying out many operations and are explained below.


Merge Click on two or more tracks and then click on this icon Merge  icon; the chosen tracks will be merged together but differentiated by colour. You can split the merged tracks by right click → *Split Merged Tracks* option.


Show density in Chromosome Selector Choose a track by clicking on it and then click on this icon Show density in Chromosome Selector  icon. For this track, its density plot will be shown alongside the chromosome.

Full view Full view icon Full View  icon will resize the image to the original size

Zoom out/ Zoom in Image zooming and zooming out can be achieved with the icons Zoom in  icon and Zoom out  icon.

Orientation By clicking this icon Orientation  icon, one can change the orientation view to make it horizontal or vertical.

Reorder Tracks The order of tracks shown in the view can be changed by clicking on the tool bar icon Reorder Tracks  icon. A window will come up listing the names, types and experiments for all the tracks. Move them around using the icons given on the right side to change the order.

Export Selected Track as Image One or more tracks can be exported as image, maintaining the order of tracks. Select the tracks for exporting and click on the icon Export Track as image  icon. Move the tracks around to maintain the order of tracks, if you choose more than one track. Define the start and end position as well as the width and height of the image required and give a file name and location for export. Apart from the image of each of the track, an html file is also saved which contains the order of the tracks.

Manage Genome Browser Data Manage Genome Browser Data  icon. See [Manage Genome Browser Data](#)

28.4.6 Track properties

On right click inside the tracks, there is a drop down which gives the following options:

- Split merged tracks, if any - If there are merged tracks, they can be split again using this option.
- Edit track properties - See section [Edit Properties](#) below for details.
- Change Track Size - Choose a track size and set the size here to resize it; height for horizontal tracks and width for vertical tracks can be changed from here.

- Export as - Gives option to export as entity list, text and image. Entity list will be added as a child under the 'All Entities' in the experiment navigator. File chooser window comes up for text and image export.
- Remove the track - Choose a track and click on this to remove the track.
- Publish - This option is enabled only in workgroup mode.

Edit Properties

'Edit Track Properties' has different options for the [Experimental Data Tracks](#) and [Annotation Tracks](#) and are explained below.

Experimental Data Track Properties

- General:**
- **Data column:** Lists all the columns available for plotting; choose one. If 'Multiple column plot' option is chosen in the 'Rendering' section, then more than one data column can be plotted in the track.
 - **Sampling:** If sampling is not applied, each data point along the visible region of the chromosome is shown in the genome browser view. This can be confusing and will also slow down the loading. To make the interpretation meaningful, **GeneSpring GX** lets the user choose to apply 'Sampling' and view either Maximum, Minimum, Mean or Median of the data points present in an interval. The interval is decided by dividing the visible region of chromosome by the display width in pixels (scaling).
 - **Smoothing:** Choose a sliding window of desired pixels (default value of 5). Within this defined window, the plot will be smoothened out with the averages values of all data points.
 - **Show Labels:** It is a very useful functionality to label the points with the actual value or the name; the actual value for each point will be shown in case of profile plots (for example, actual copy numbers will be shown for copy number experiments in profile plot); In case of region plot, one can label by any datatype with the 'Label by' option present under 'Show labels'.
 - **Ruler:** A linear ruler will divide the region above the reference and below the reference into equally spaced areas. If 'Linear ruler' option is unchecked, the plot will be non-linear; the regions will be divided so as to show a maximum zoom in the area just above and below the reference and this area would progressively decrease as you move away from the reference line.
 - **Auto Calibrate:** For all plots other than region plot, there is an option to auto calibrate which automatically bins the region between min and max set in the ruler.
 - **Reference, Minimum and maximum values, Number of lines** The minimum and maximum values set the boundaries for the Y axis. Between the reference and the minimum value (as well as between the reference and the maximum value), the region is horizontally divided into areas based on the number of lines defined under the ruler option.
- Rendering:** Gives option to choose the type of plot to be shown. For expression experiments, a scatter plot is shown by default, while it is a profile plot for copy number experiments. Region plot shows the region of the genome in the visible pixel. Multiple column plot allows visualizing more than one data column in the plot.

Selection: Allows setting a border width for each track and also to define a colour to show the selected track (or the track set as navigation track). The entire track will be coloured if background colour option is checked.

Annotation Track:

Annotation tracks are shown as region plots. Table 28.1 describes some common annotation tracks.

Table 28.1: Annotation Track Properties

Transcript Track	One can visualize gene, transcript and exon region in the genome browser.
Gene Track	One can visualize gene.
CpGIIsland Track	Shows the CpGIIslands on the genome; user can choose to view any of these as data columns - Length, CpG Number, CG number, CpG percent, GC percent, Observed Exp .

28.5 Viewing Copy Number Experiments in Genome Browser

28.5.1 Data columns

For Copy number experiments, the following data can be chosen from the experiment to be viewed in the Genome Browser:

1. Copy Number
2. Copy Number Confidence
3. LOH Score (Not supported for Illumina)
4. Log Ratio
5. Allele Specific Copy Number
6. Parent Specific Copy Number

In addition to the above, Mean Log Ratio from CBS can also be viewed as a data column. See section [Terminology in Copy Number Analysis](#) for a description of these values.

28.5.2 Utilities for Copy Number Experiments

- Highlight all tracks - From the spreadsheet, select any row and a track; the position of the SNP will be indicated in the track; Right click in spreadsheet and say 'Highlight all tracks'; it will be shown in all tracks.
- Exported segments can be viewed in Genome browser. Drag and drop the exported segments into the browser and go through the steps explained in section [Drag and Drop Files from anywhere](#).
- The chromosome start and end position are same for SNP and CN probes. For CN probes, it is taken as $(\text{Start}+\text{End})/2$.
- For [PSCN](#), the 'diff' values are shown by default; User can right click →Edit Track properties →Multiple column plots and check all the three data - min, max and diff values to be shown in the same track.


NOTE: For Chromosome Y, [Birdseed algorithm](#) does not generate calls and clusters for more than 60% of the SNP probes; and so Copy Number Analysis does not run for SNP probes for chromosome Y. Hence SNP probes are not seen when Chromosome Y is viewed in the Genome Browser.

28.6 Useful details to know

1. As long as there is chromosome name, start and end location, this information can be viewed inside the genome browser along with any other associated data.
2. Selections are honoured across the various views inside the genome browser.
3. Double click on any entry in the spreadsheet to zoom the region in the track.
4. Any file imported using **Drag and Drop** can also be exported as text.
5. Tracks are added progressively into the Genome Browser; there is an option to cancel this process from the Progress Bar. If the memory exceeds the limit, then the process would be truncated.
6. When genome browser is launched, it shows by default the builds of the organism in that active experiment.
7. Though genome browser is not associated with any particular experiment in **GeneSpring GX** , it can be launched only if there is an active (open) experiment. Hence, while dragging and dropping files from outside the tool also, one needs to have an active experiment and then launch the genome browser and populate it with files. These files are saved as part of annotations and are not associated with any experiment inside **GeneSpring GX** .

28.7 FAQ

1. **How can I see the details about the organism/build/feature relevant to the tracks?**

Within the Genome Browser View, click on the icon Manage Genome Browser Data  icon. A window will come up with an explorer on the left side, with a tree like structure of existing organisms/builds/features. Expand the tree and double click on any item to views details about it on the right side of the window.

2. **Why and when should I apply 'Sampling'?**

If 'Sampling' is not applied, each data point along the entire length of the chromosome is shown in the view. This can be confusing, especially in the context of copy number and association experiments where the number of data points for each sample are in the order of millions. For meaningful interpretation with large data, user can choose to view either Maximum, Minimum, Mean or Median of the data points present in an interval. The interval is decided by dividing the visible region of the chromosome length by the display width in pixels (scaling).

3. **What is the difference between 'Sampling' and 'Smoothing'?**

In the case of sampling, the visible region of the chromosome length is divided by the display width in pixels (scaling) and data is shown here obeying the type chosen (maximum, minimum, mean or median). While in the case of smoothing, the user can define a window within which all the data points will be averaged.

4. **What is a region plot?**

Plot showing a region of the genome ; region plots can be labelled by the data column. Region plot is the default plot for entity lists and annotation tracks.

5. **How do I set the Reference?**

Reference enables comparison against a standard. **GeneSpring GX** has set intelligent defaults for reference, meaningful for comparisons, but the user has the option to change the reference using the 'Edit Track Properties →General' option.

6. **My plot does not look like what I expected. What could be the reasons?**

- (a) Check the set reference value and reset it.
- (b) Try applying 'sampling' or 'smoothing'.
- (c) Try with Linear and Non linear ruler with different reference points.
- (d) Change the type of plot. For example, a profile plot with filled colour is a better view for copy number experiments while a region plot with 'Colour by a list associated value' may be more meaningful for entity lists.

7. **From my expression experiment, I dropped multiple samples into the genome browser but I do not see all them. Why?**

If you chose to view data from an averaged interpretation, then the samples are averaged on the basis of the condition and only the condition is shown in the genome browser. This is true for expression experiments. If you want to see each sample separately, use an unaveraged interpretation.

8. **I want to add an annotation track, say CpGIIsland track to an existing build. How do I do this?**
Click on 'Manage Genome Browser Data' icon, go to an existing build and click on 'Add feature'. Do an advanced import with the CpGIIsland file. Now, you should be able to see the CpGIIslands in the genome browser.
9. **Can I use any annotation file like transcript file to add a build?**
Only cytoband files can be used to add a build as they span the entire genome.
10. **I have an active experiment. When I try to launch Genome Browser, it shows the message 'Organism is not supported'. What should I do?**
GeneSpring 11.0 comes prepackaged with annotation data for humans (Hg 18 and Hg19), mouse and rat obtained from UCSC <http://hgdownload.cse.ucsc.edu/downloads.html>. Go to the menu **Annotation** → **Update Genome Browser Data** → **From Agilent Server (or GeneSpring Update File)** and download the annotation data for the required organism. If the organism in the active experiment is not 'Homo Sapiens' (humans) or 'Mus Musculus' (Mouse) and 'Rattus Norvegicus'(Rat), then User can add this organism by using '*Manage Genome Browser Data*' from the menu **Annotation** → **Update Genome Browser Data** → **Manage Genome Browser Data**' explained in section [Manage Genome Browser Data](#).
11. **How can I see the actual copy numbers in the tracks?**
While dragging and dropping experiments containing copy number values, right click on the track, edit track properties, and click 'Show Labels' under General tab. If copy number is one of the data column in the file viewed in Genome Browser, then right click on the track, edit track properties, and click 'Show Labels' under General tab and choose the 'Copy Number' under data column. The values will be shown.
12. **When I plot Copy Number and Allele Specific Copy Number (ASCN), the points do not match. Why?**
For V6 experiments, the copy number is computed for all SNP and CN probes while the ASCN is computed only for SNP probes and hence there are fewer points in ASCN track. PSCN track will also show only SNP probes.

Chapter 29

Ingenuity Pathways Analysis (IPA) Connector

The **GeneSpring GX** -IPA Connector enables users of **GeneSpring GX** and Ingenuity Pathways Analysis (IPA) to exchange information seamlessly between the two applications. Genes of interest can be identified using powerful statistical and analytical tools in **GeneSpring GX** . The biological context and significance of these findings can then be assessed in IPA using various tools. With the **GeneSpring GX** -IPA Connector, users will be able to send Gene Lists and associated expression data directly to IPA. Once in IPA, users can perform network analyses, build pathways, view relevant canonical pathways, and obtain proprietary information on protein interactions and pathways. Users can then send a list of genes back to **GeneSpring GX** , allowing further iterative analysis of those genes in **GeneSpring GX** .

These options can be accessed from *Results Interpretations*.

29.1 Using the GeneSpring GX -IPA Connector

29.1.1 Create Pathway in IPA

The Create Pathway in IPA option will allow users to send an Entity List from **GeneSpring GX** to IPA and use those genes to create a pathway in IPA. This pathway can then be subjected to further manipulation and analysis in IPA by growing a node, removing nodes and interactions, and interrogating a node or an interaction. Users will be able to create gene lists from selected genes in the pathway and send the gene lists back to **GeneSpring GX** for further analysis. To avail this option, select *Launch IPA* from the *Results Interpretations* section in the Workflow browser. See figure [29.25](#).

The IPA dialog box appears with the three activities that are supported in **GeneSpring GX** . Select *Create Pathway in IPA*. See figure [29.2](#).

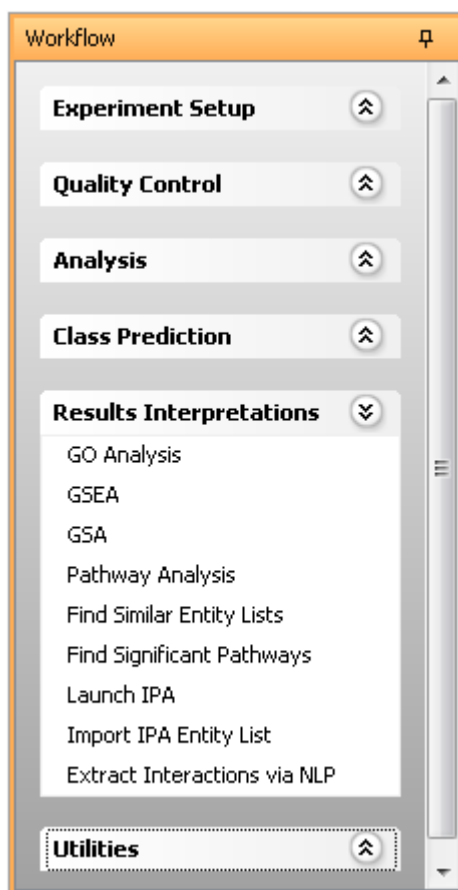


Figure 29.1: Launch IPA

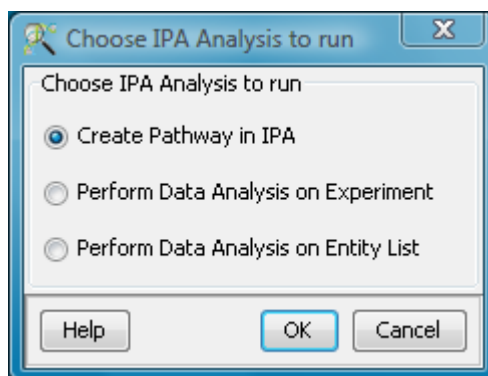


Figure 29.2: Create Pathway in IPA

The *Create New Pathway* dialog box (29.3) will show up:

- **Entity List:** By default, the active entity list is selected. Press 'Choose' to select a different Entity List.
- **IPA Server address:** Name of the server running IPA. By default this will point to the main server of Ingenuity "analysis.ingenuity.com". To choose a different server, enter the server address of desired server. IPA Server address can also be permanently configured from *Tools*→*Options*→*Miscellaneous*→*Pathway Analysis(IPA)*
- **Pathway Name:** The name for the pathway. By default, the name of the entity list that was originally selected will be used. If you selected a different Entity List in this dialog the name for the pathway will not be updated to reflect the new selection.
- **Project Folder:** The name of the project in IPA this pathway should be stored under. By default, it will use the name of the **GeneSpring GX** project.
- **Gene Identifier Column:** This indicates the type of gene identifier that will be used to map genes in the Entity List to genes in the Ingenuity Pathways Knowledge Base (IPKB). The type of identifier selected, determines which annotation column in the **GeneSpring GX** technology the identifiers are retrieved from. The list of supported identifiers is given below. These identifiers will then be used to map genes in the list to genes in IPKB. Only identifiers that can be matched to genes in IPKB will be used to build a new pathway in IPA.

Identifiers:

- Entrez Gene ID
 - Locus Link ID
 - Affymetrix Probeset ID
 - UniGene ID
 - GenBank Accession
 - Swissprot
 - Agilent Probe ID
 - RefSeq Protein ID
- **Save Pathway:** Indicates whether the pathway generated in IPA will be saved. If selected, pathway generated will be saved in IPA to the specified Project Folder, within My Pathways, under the specified Pathway Name. Press **OK** to create the pathway in IPA. Your default browser will start up and connect to the IPA server.

IPA Pathway Creation

The connection to IPA is performed through the IPA HTTP API and on some systems this may result in the session being blocked due to the security settings as shown below. If this happens, simply click on the

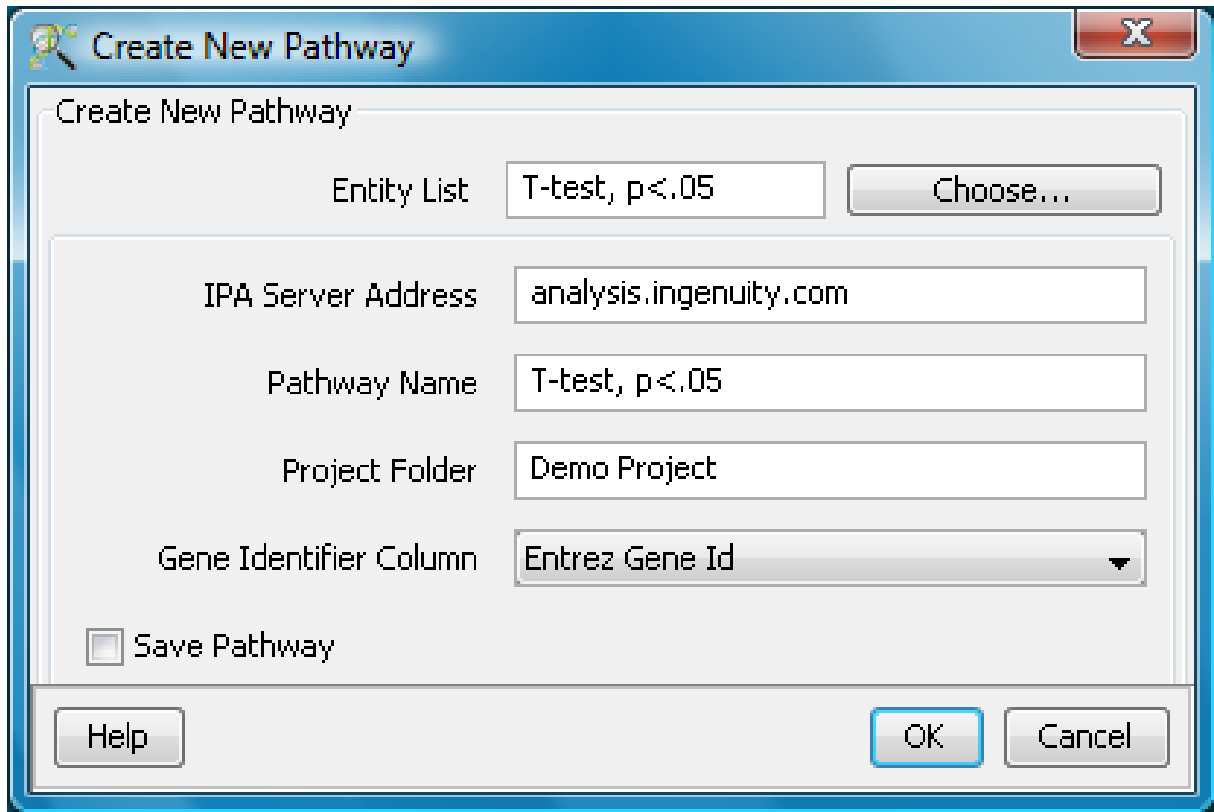


Figure 29.3: Create New Pathway

yellow bar and select 'Allow Blocked Content'. There is no danger in allowing this access. See figure 29.28

IPA is using Java Webstart and a dialog appears indicating JAVA is starting up. See figure 29.29.

The login dialog will appear. Enter your login credentials here. See figure 29.30

IPA will start up and show the pathway as shown in figure 29.7

At this point you can continue your analysis in IPA.

29.1.2 Import List from IPA

The user can send a list of entities from IPA back to **GeneSpring GX** for further analysis in **GeneSpring GX** , create a list in IPA and use the *Import IPA Entity List* in the *Results Interpretations* section. To create a list of pathway genes in IPA:

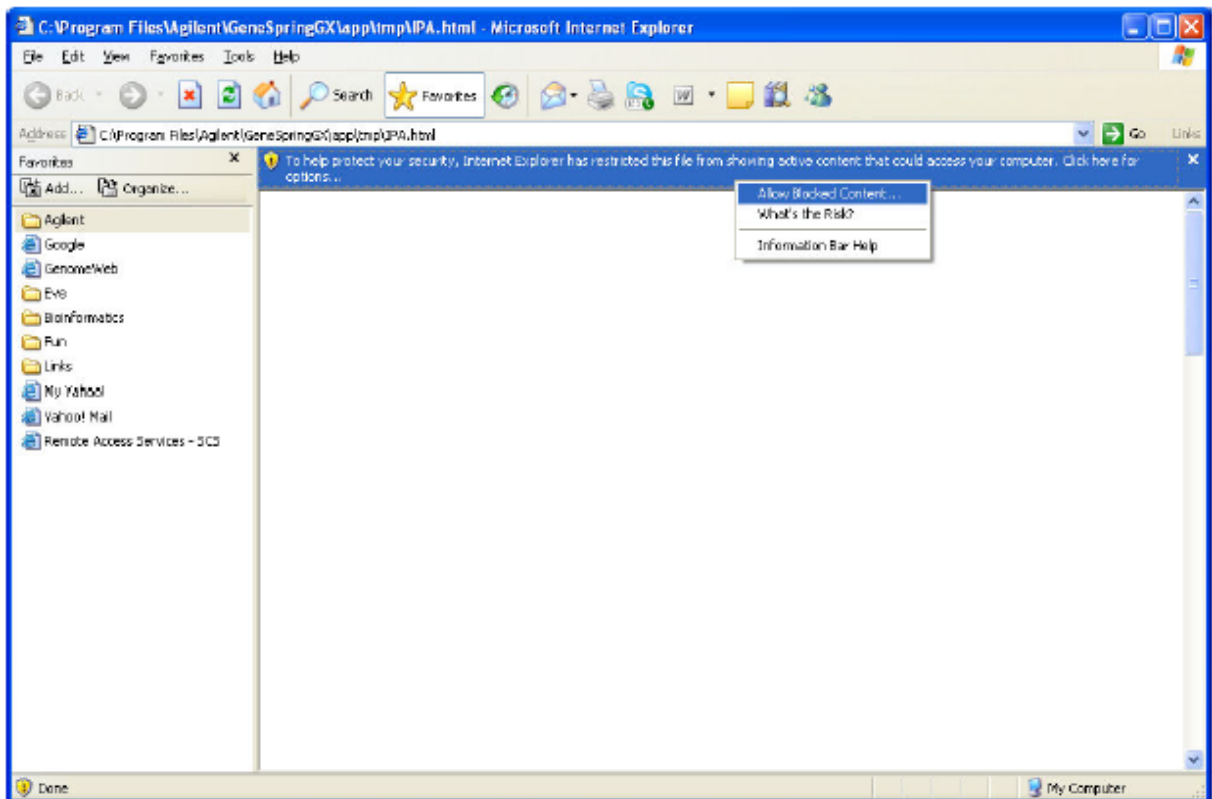


Figure 29.4: IPA Pathway Creation



Figure 29.5: Java Startup



Figure 29.6: IPA Login Dialog

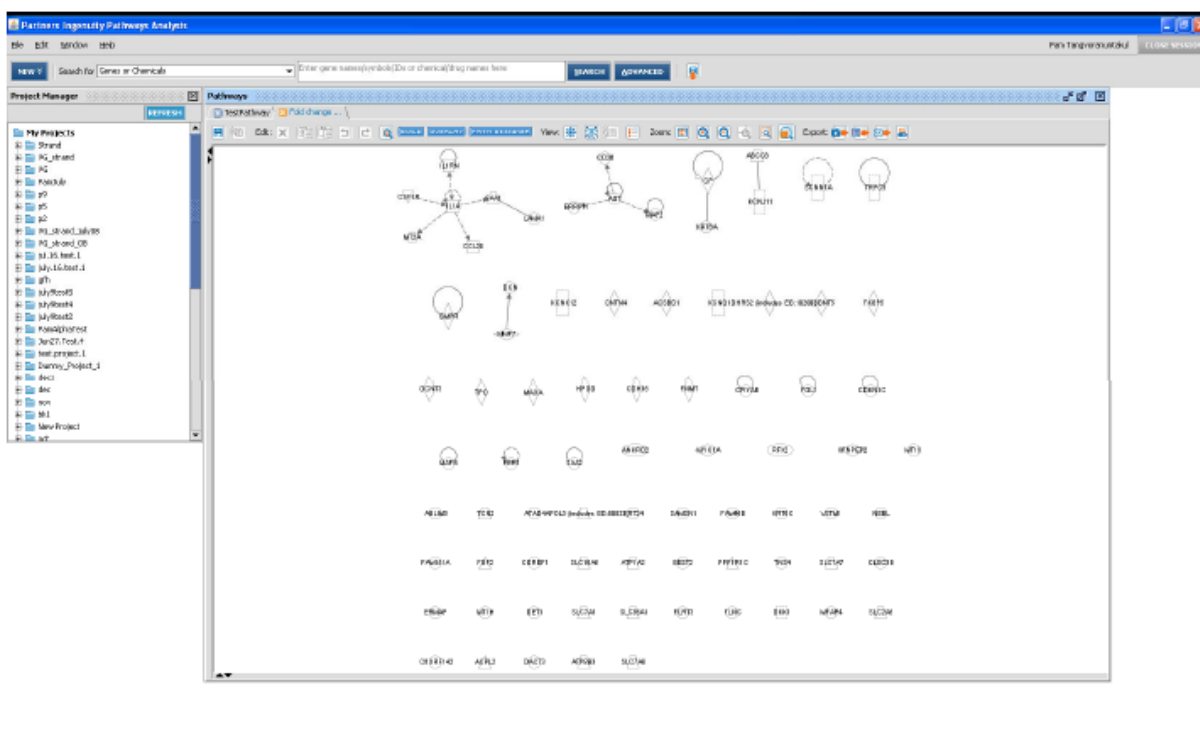


Figure 29.7: Pathway Analysis in IPA

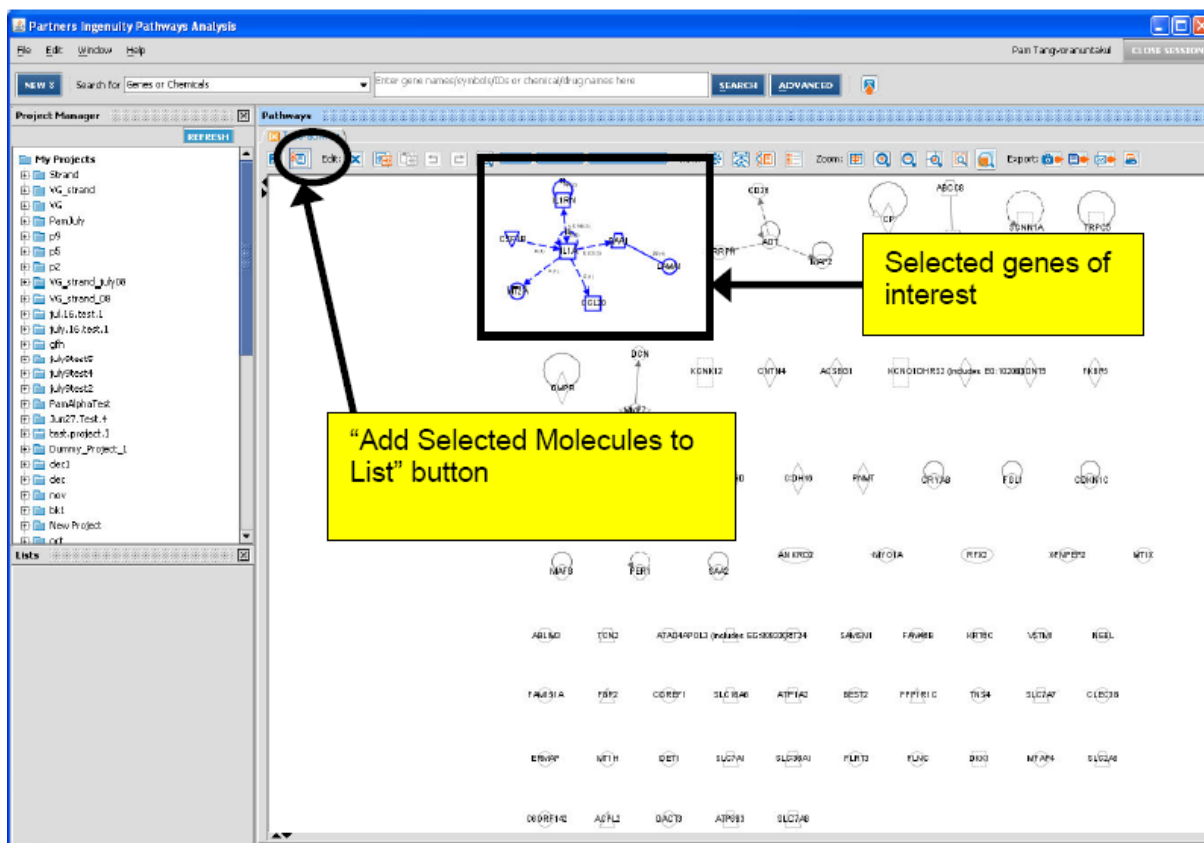


Figure 29.8: Creation of Entity List

1. Select the molecules of interest by dragging and drawing a box around the molecules of interest in the pathway view. These molecules will then be highlighted. Click on the 'Add Selected Molecules to List' button and select 'New List' to create a new list or 'Saved List' to add the selected genes to an existing gene list. See figure 29.8

Alternatively, select the genes for export to **GeneSpring GX** in the Network tab of the analysis results and create a list from them. See figure 29.9

2. Indicate List Name and Project to save the gene list under. The gene list will be saved in the specified Project folder under My Lists. Selected genes of interest 'Add Selected Molecules to List' button. See figure 29.10
3. Click on the black arrow. Select 'Send to' and select **GeneSpring GX** . Specify the settings such as the type of identifier to use to map genes in the gene list to genes in the **GeneSpring GX** genome. Click **OK**. See figure 29.11
4. Indicate the location to where the file will be saved. The gene list will be saved as a **GeneSpring GX** zip file. See figure 29.12

To import the gene list into **GeneSpring GX** , use the option 'Import List from IPA', described below.

Differentially expressed - 2006-03-17 01:27 PM

Networks \ Functions \ Canonical Pathways \ My Pathways \ Gene Summary \ Network Explorer \ Overlapping Networks

FILTER [] VIEW NETWORKS ADD TO PATHWAY ADD TO LIST MERGE NETWORKS >>

The analysis is composed of 8 networks. To view a network, select the appropriate network(s) and click the View Networks button. To merge selected networks, click Merge Networks.
Total selected nodes: 35

<input type="checkbox"/>	^	ID	Genes	Score	Focus Genes	Top Functions
<input checked="" type="checkbox"/>		1	BAG1, BCL2, BNIP3L, CLCN3, CSF2RA, DLL1, DOK2, EMR1, ENO2, GNA13, HBB, HBE1*, HNRPA1, HOXC8, HRK, IL3, ITGAV, ITPR1, JAG1, KLF1, MSH42, NFATC3, NOTCH1, NOV, PDK3R2, RALA, RALBP1, RBPSUH, REPS1, RLBP1, Serpina3g, SERPINF1, SPP1, TEC, USF2	23	15	Cancer, Cellular Development, Cellular Growth and Proliferation
<input type="checkbox"/>		2	ARF6, ASNS, ATF4, C12orf14, CCL6, CEND3, Crisp1 (MGI:102553), DOK2, E2F1, EIF251, GJA1, GUSB, HOXB4, HRAS, HSPA1A, KSR, Mepe2, MGST3, MTLA, MYC, NCAM1, NOTCH1, NSEP1, PRDX2, PRTN3, PSMB9, RASA4, RPS6, SERPINB2, SOX4, TFDP1, Tgtp, TIE1, TNF, TRA1	23	15	Cell Cycle, Cancer, Cellular Growth and Proliferation
<input type="checkbox"/>		3	ACPI, ACP5, ALOX5, ALOX5AP, BHLHE2, C3, CALCA, CCL11, CCR3, CD14, CDH2, CPOX, CTLA4, FAAH, FCER1A, FLNB, GNA15, IL4, IQGAP1, ITGAX, ITGB1, ITGB7, JAK3, LSP1*, LTA, MAPK1, NCAM1, PFC, S100A11, SPP1, SPTAN1, TGFBI, THY1, TRG@, XBP1	21	14	Cellular Movement, Hematological System Development and Immune Response
<input type="checkbox"/>		4	ATM, CSF1R, CTNNB1, DDIT3, DIO2, EGFR, EIF2AK2, F7, FGF3, FRAT2, GADD45A, GADD45B, GSK3B, HBEGF, HEMGN, HMOX1, HNF4A, KRT5, LEF1, LMNA, MAPT, MATN2, PACSIN1, PERP, PLAT, PPF2CA, PPP2R1A, SERPINB2, SERPINE1, SLC2A1, SMAD4, SPP1, STK11, TMSB10, TP53	19	13	Cellular Growth and Proliferation, Cell Death, Cancer
<input type="checkbox"/>		5	AHCY, ALAD, C10orf58, CA2, CCL4, CCL6, CD14, CEBPE, CISH, DDIT3, DOK1, EIF4EBP1, FKBP1A, FRAP1, HCA112, IGFBP2, IGFBP4, IL5, IL13, IL1RN, IL3RA, Ins1, KITLG, LR8, MMP1, MMP14, NFATC3, PPARA, RHAG, RPS6, RXRA, SNAP23, SOD2, TERT, TNFSF4	12	9	Cell-To-Cell Signaling and Interaction, Hematological System Development and Inflammatory Disease

Figure 29.9: Creation of Entity List

Select the **Import IPA Entity List** link in the **Results Interpretations** section in the workflow browser. See figure 29.25.

This would prompt you with a dialog to select the zip file created from IPA in the previous step. Select the zip file and press **Open**. See figure 29.14

The option will import the gene list stored in this zip file to a new Entity List with the same name the user provided in the previous step in IPA. The new Entity List will be saved in a folder called 'Imported Lists' (it would be created automatically if necessary). See figure 29.15

Note: Only identifiers listed in the previous section shall be supported in matching the imported genes to entities in the current technology. If no entities can be matched, then the option would show an error as shown in figure 29.16

Select a different identifier in IPA to represent the genes. It is recommended to choose EntrezGene IDs since many of the technologies in **GeneSpring GX** will have these identifiers.

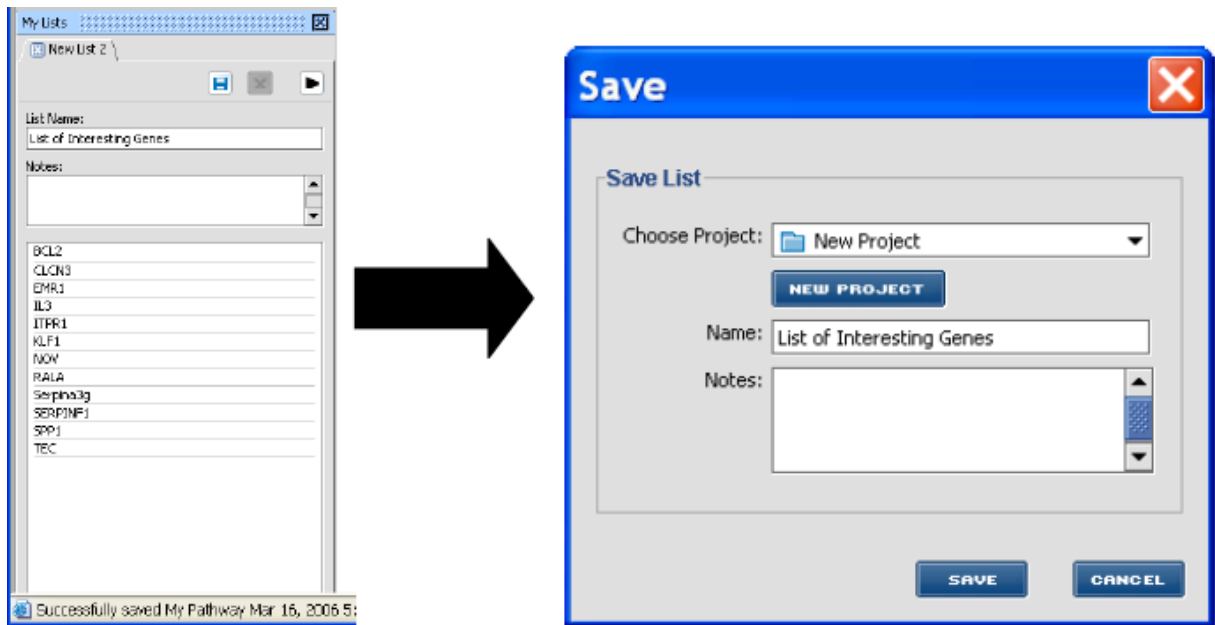


Figure 29.10: Save List

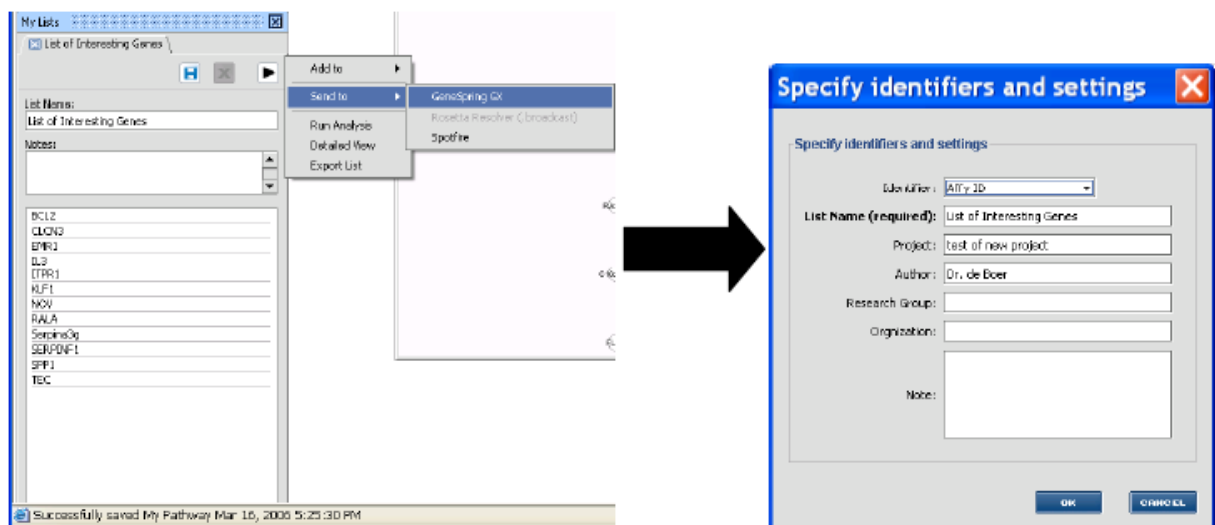


Figure 29.11: GeneSpring GX suitable list creation

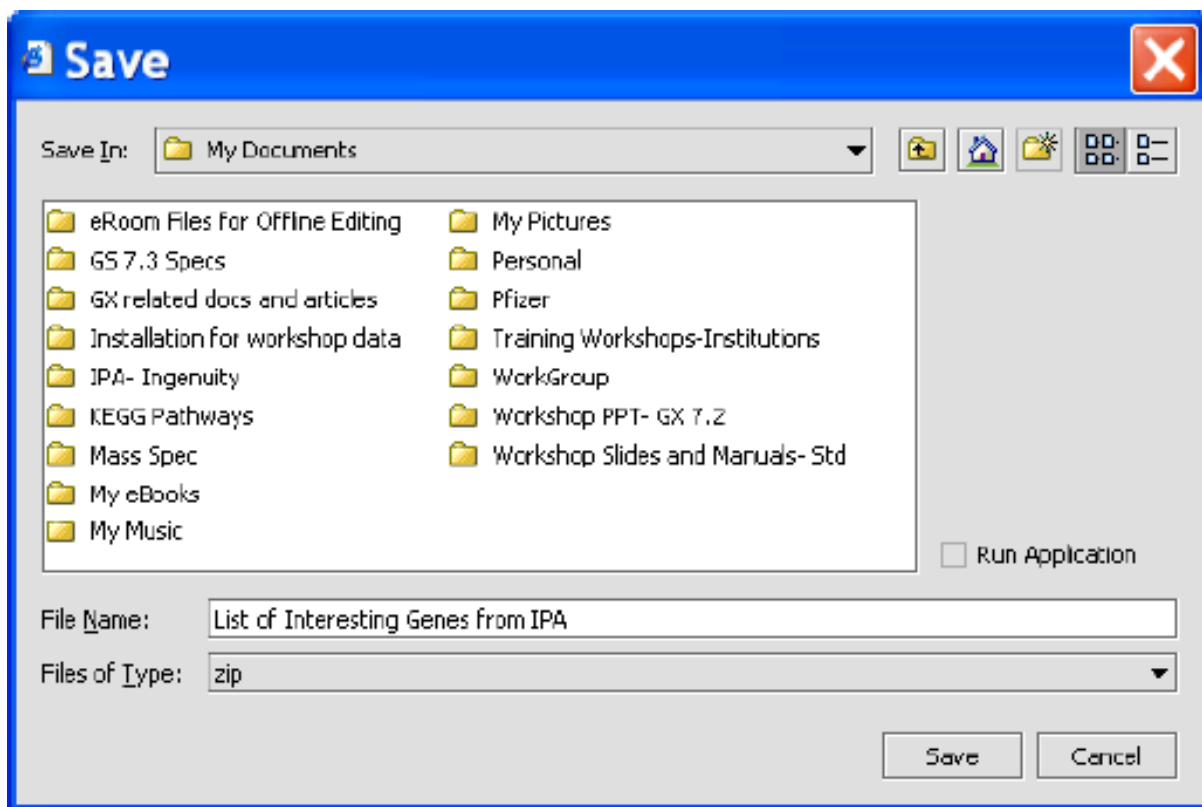


Figure 29.12: Saved List Location

29.1.3 Perform Data Analysis on Experiment

This option will allow you to send an Entity List and the associated gene expression data from **GeneSpring GX** to IPA to perform data analysis in IPA. Genes on the Entity List that are also found in IPKB will be used as Focus Genes to build networks. The networks can be subjected to further manipulation and analysis in IPA by growing a node, removing nodes and interactions, interrogating a node or an interaction, and perform Function, Canonical Pathways, My Pathways, Gene Summary, and Overlapping Networks analyses. Users will be able to create gene lists from the generated networks and send the gene lists back to **GeneSpring GX**.

To utilize this option, select *Launch IPA* From the *Results Interpretations* section in the Workflow browser. See figure 29.25.

The IPA dialog box appears with the three activities that are supported in **GeneSpring GX**. See figure 29.18

Select 'Perform Data Analysis on Experiment' and the dialog box below (29.19) will show:

Inputs:

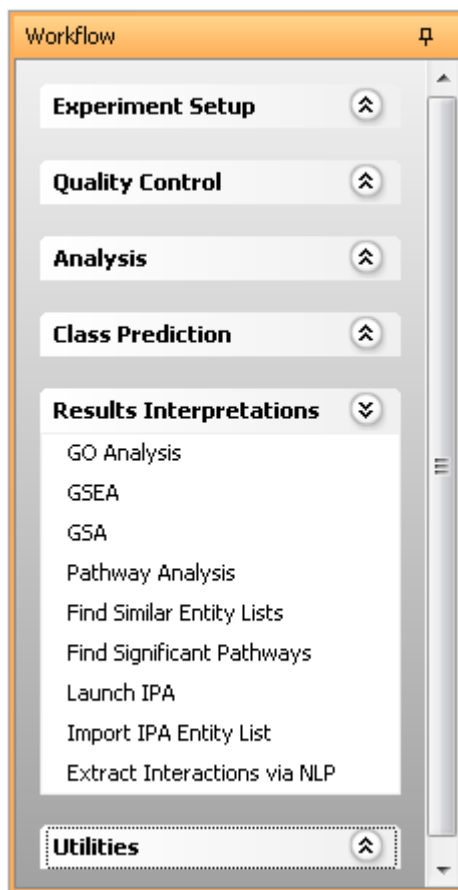


Figure 29.13: Import IPA Entity List

- **Entity List:** The active entity list is selected. To use a different entity list, cancel the option, select a different entity list and select the option again.
- **Experiment Interpretation:** The active Interpretation is selection. To change the interpretation, press 'Choose' to change the interpretation. The experiment interpretation will define the conditions for which the log2 values for each gene will be sent to IPA. Log2 values for the conditions in the selected experiment interpretation will be sent to IPA for analysis. The name of the data set used in IPA will be the name of the experiment. NOTE: IPA only allows unique names for datasets per project. To analyze the same experiment more than once, either change the name of the experiment or change the project name.
- **IPA Server address:** Name of the server running IPA. By default this will point to the main server of Ingenuity 'analysis.ingenuity.com'. To choose a different server, enter the server address of desired server. IPA Server address can also be permanently configured from *Tools* → *Options* → *Miscellaneous* → *Pathway Analysis(IPA)*
- **Project Folder:** The name of the project in IPA this pathway should be stored under. By default, it will use the name of the **GeneSpring GX** project.
- **Use both Direct and Indirect relationships:** If selection is 'yes', IPA will build networks using

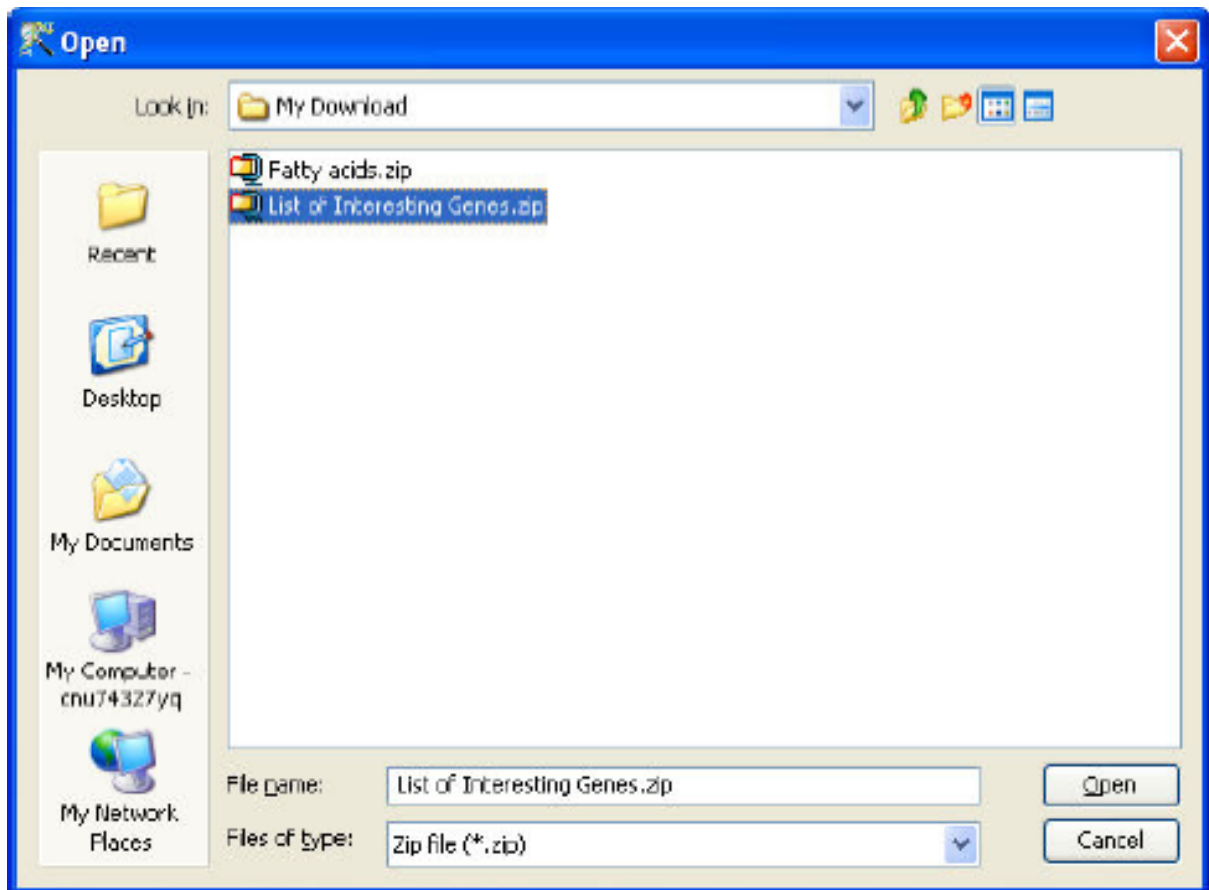


Figure 29.14: Selection of Folder

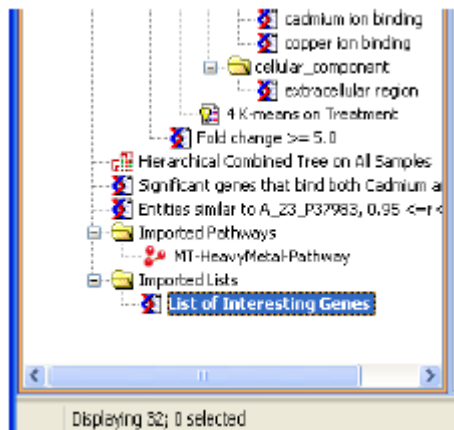


Figure 29.15: Entity List Creation

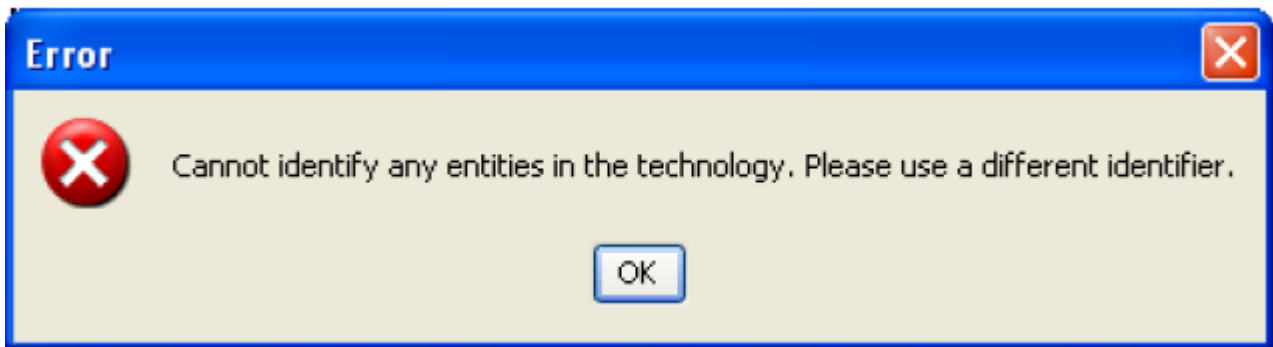


Figure 29.16: Error Message

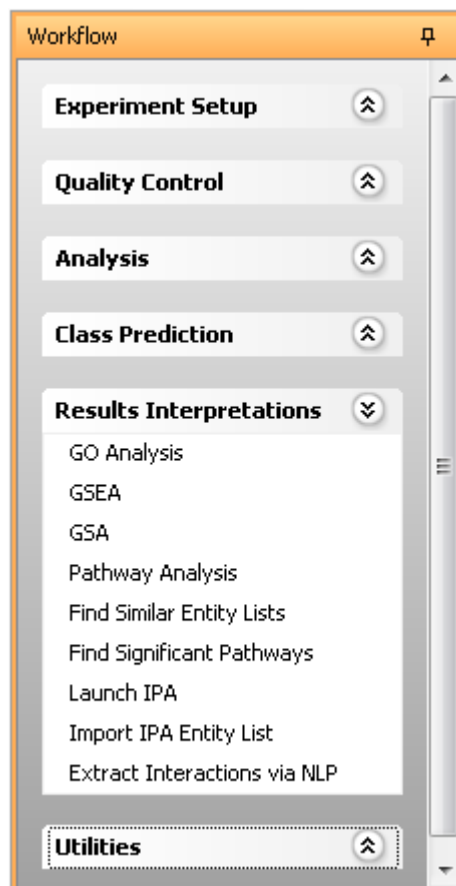


Figure 29.17: Launch IPA

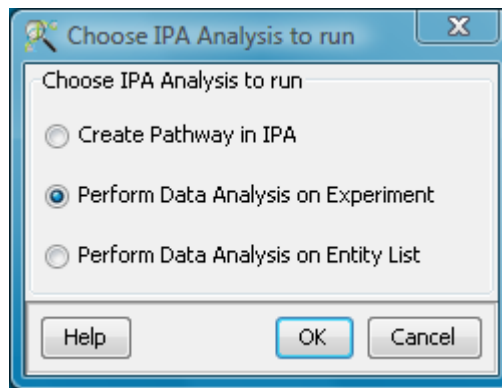


Figure 29.18: Data Analysis on Experiment

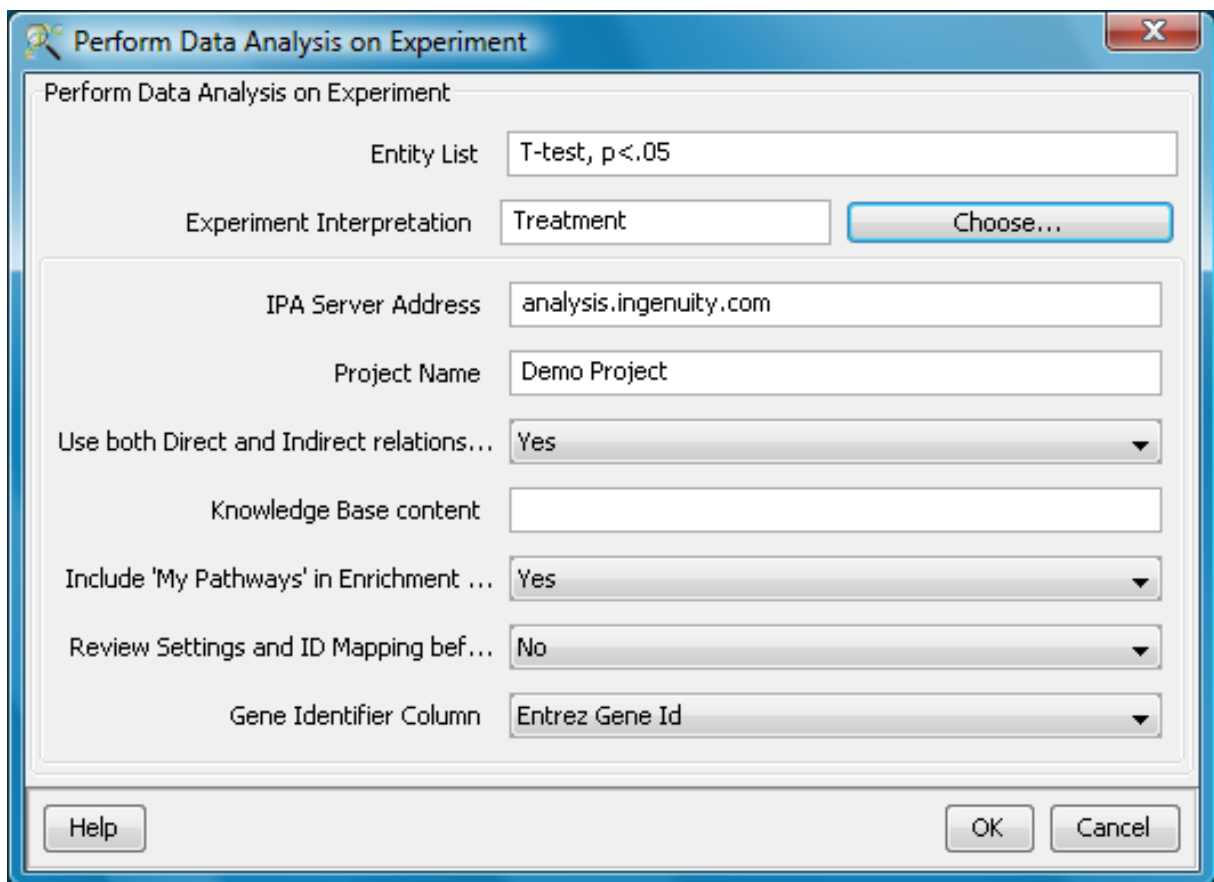


Figure 29.19: Perform Data Analysis on Experiment

both direct and indirect molecular interactions between genes. If selection is 'no', IPA will build networks using only direct interactions between genes.

- **Knowledge Base content:** This indicates what database will be searched for information to build the network. An empty string indicates all available Knowledge Bases will be searched and information from all sources will be used to perform analysis.
- **Include My Pathways in Enrichment Score:** If selection is 'yes', all pathways saved under My Pathways will be included in the scoring process.
- **Review Settings and ID Mapping before Running Analysis:** If selection is 'no', data analysis will be automatically performed using the settings defined in this option. If selection is 'yes', the Create Analysis window below will be displayed for users to review and modify settings before running analysis.
- **Gene Identifier Column:** This indicates the type of gene identifier that will be used to map genes in the Entity List to genes in the Ingenuity Pathways Knowledge Base (IPKB). The type of identifier selected, determines which annotation column in the **GeneSpring GX** technology the identifiers are retrieved from. The list of supported identifiers is given below. These identifiers will then be used to map genes in the list to genes in IPKB. Only identifiers that can be matched to genes in IPKB will be used to build a new pathway in IPA.
 - Entrez Gene ID
 - Locus Link ID
 - Affymetrix Probeset ID
 - UniGene ID
 - GenBank Accession
 - Swissprot
 - Agilent Probe ID
 - RefSeq Protein ID

Press 'OK' to start the analysis in IPA. Your default browser will start up and connect to the IPA server.

IPA analysis

The connection to IPA is performed through the IPA HTTP API and on some systems this may result in the session being blocked due to the security settings as shown below. If this happens, simply click on the yellow bar and select 'Allow Blocked Content'. There is no danger in allowing this access. See figure [29.28](#)

IPA is using Java Webstart and a dialog appears indicating JAVA is starting up. See figure [29.29](#).

The login dialog will appear. Enter your login credentials here. See figure [29.30](#)

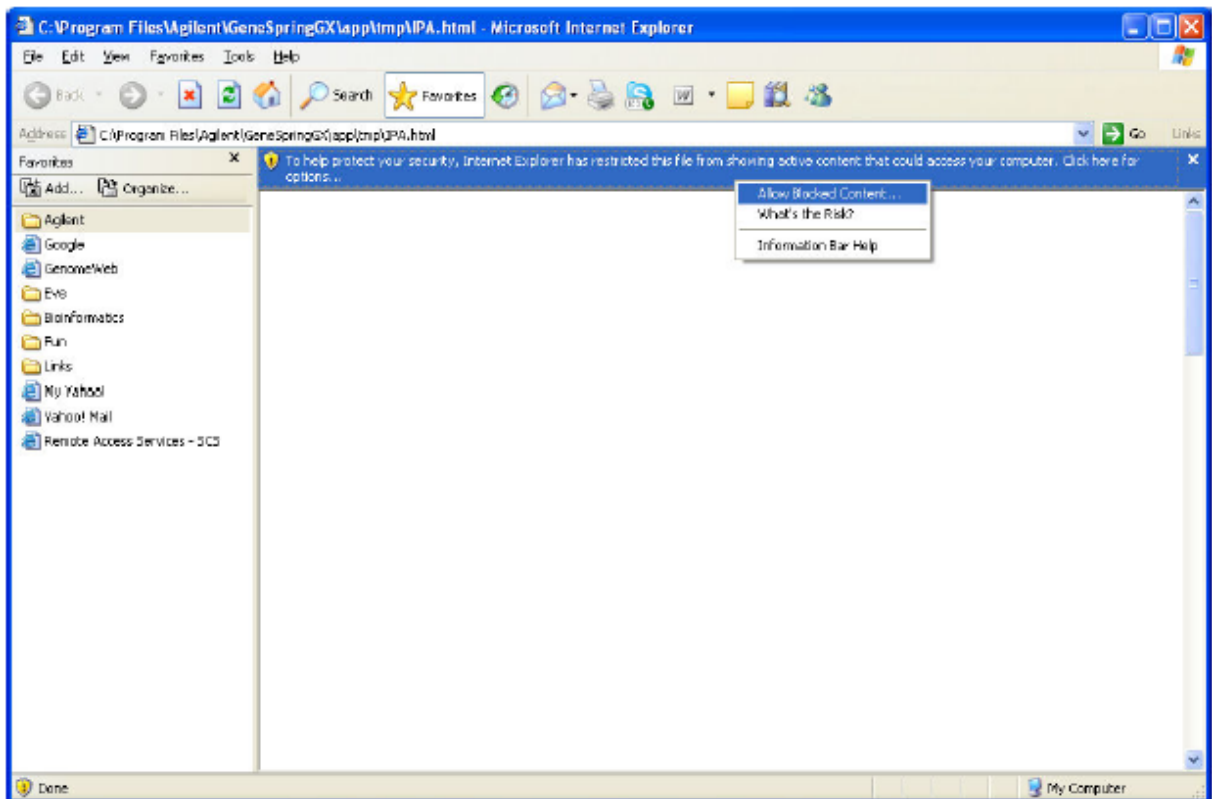


Figure 29.20: IPA Pathway Creation



Figure 29.21: Java Startup



Figure 29.22: IPA Login Dialog

If 'Review settings and ID mapping before running an analysis' was selected, the following screen 29.31 in IPA will appear, if not, the main window will show.

Review the analysis settings and press 'Run Analysis' to proceed. See figure 29.32

The analysis is performed and the Analysis will have a little clock in the icon while it is running. When the analysis is finished, the clock will disappear the analysis is ready for review.

Expression values are assumed to be Log Ratio

The expression values that are sent to IPA are assumed to be Log Ratios (positive and negative values centered around 0, in log 2 space). Most experiments in **GeneSpring GX** are using Log Ratio values, due to the baseline transformations of the experiment that are typically used in the experimental designs. This assumption may not always hold true for your experimental design.

29.1.4 Perform Data Analysis on Entity List

This will allow you to send a Entity List, with or without list associated values, from **GeneSpring GX** to IPA to perform data analysis in IPA. Genes on the Gene List that are also found in IPKB will be used as Focus Genes to build networks. The networks can be subjected to further manipulation and analysis in IPA by growing a node, removing nodes and interactions, interrogating a node or an interaction, and perform Function, Canonical Pathways, My Pathways, Gene Summary, and Overlapping Networks analyses. Users

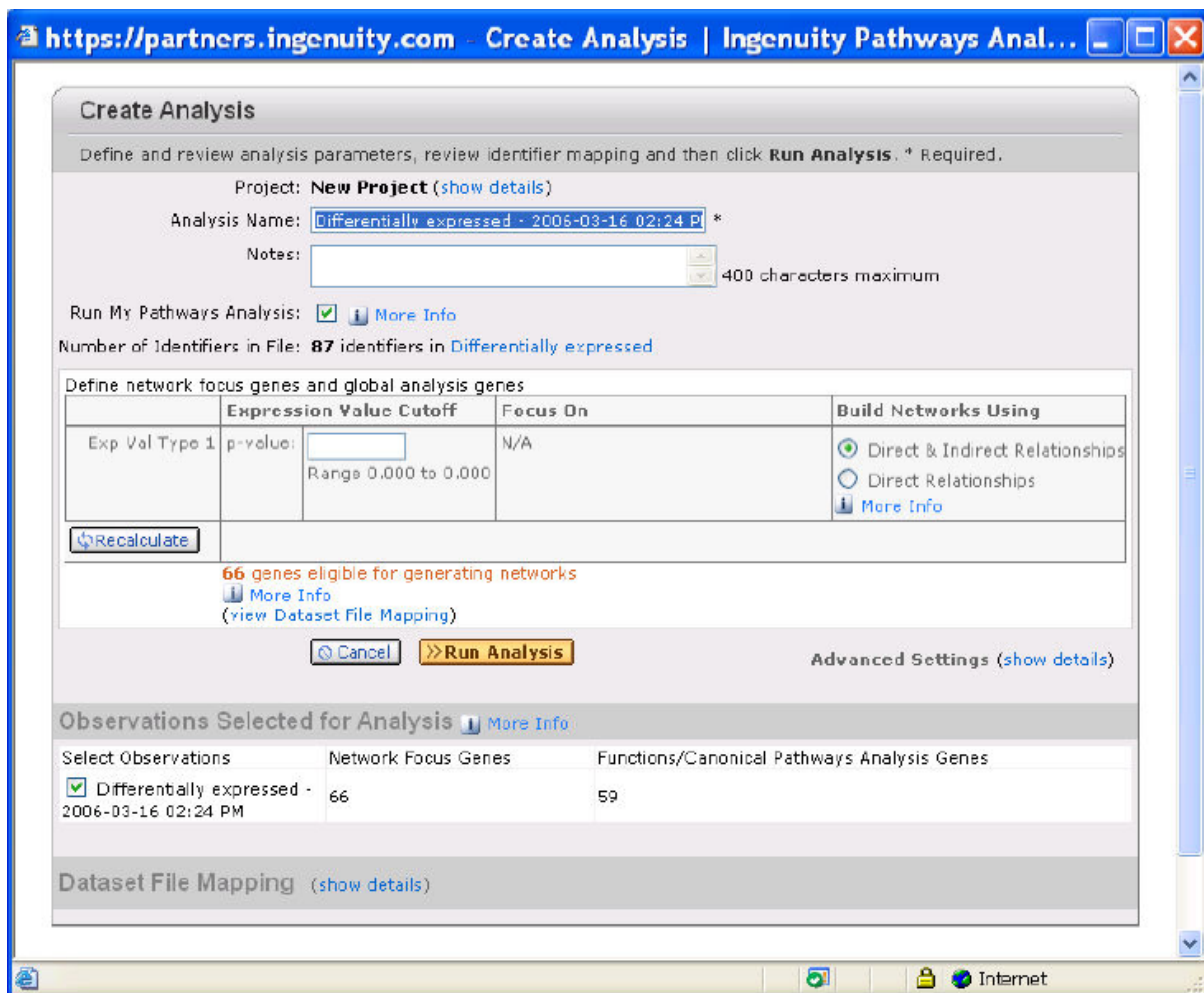


Figure 29.23: Create Analysis

will be able to create gene lists from the generated networks and send the gene lists back to **GeneSpring GX** .

To utilize this option, select *Launch IPA* from the *Results Interpretations* section in the Workflow browser. See figure 29.25.

The IPA dialog box appears with the three activities that are supported in **GeneSpring GX** . See figure 29.26.

Choose *Perform Data Analysis on Entity List* and the options dialog box (29.27) will appear:

Inputs:

- **Entity List:** The selected entity list is selected. To use a different entity list, cancel the option,

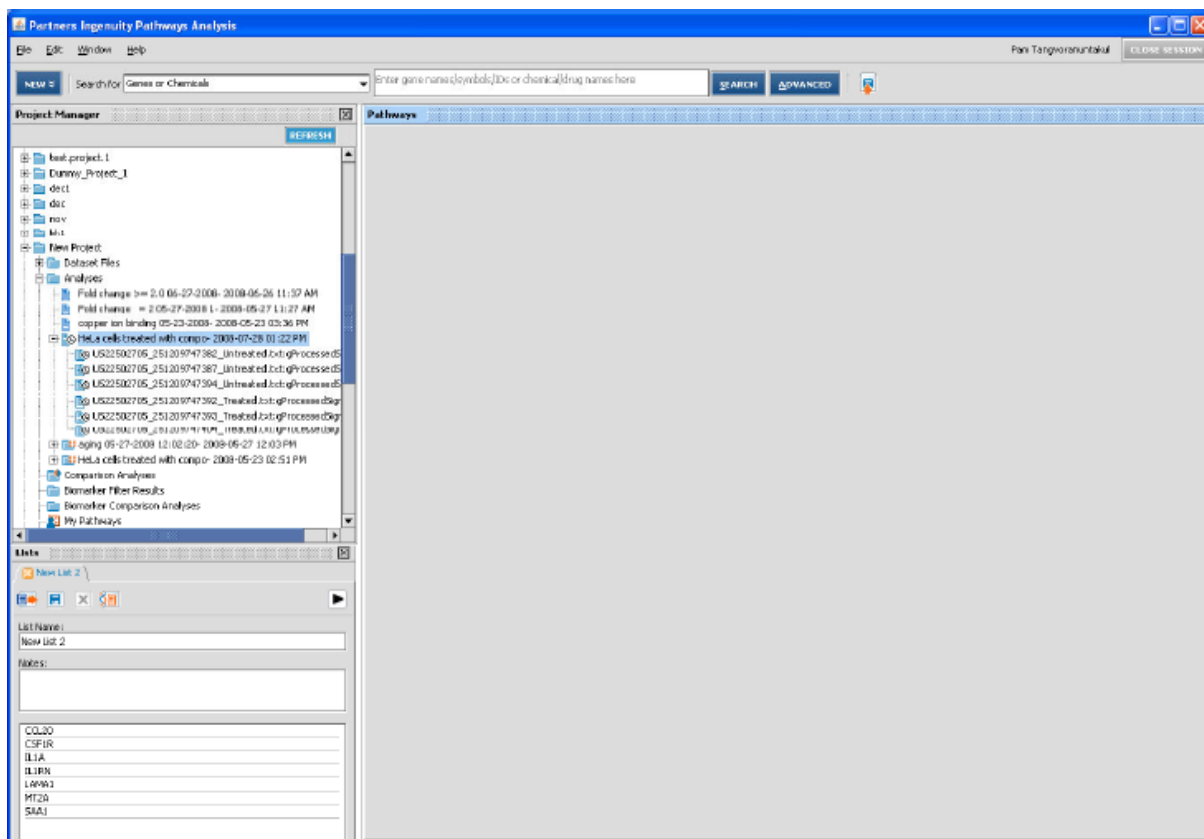


Figure 29.24: Analysis Settings

select a different entity list and select the option again.

NOTE: IPA will use the name of the Entity List to name the data set in IPA. IPA will only allow unique names for the datasets per project. To perform a different analysis on the same entity list, perform the analysis from within IPA or change the project name in this dialog.

- **IPA Server address:** Name of the server running IPA. By default this will point to the main server of Ingenuity 'analysis.ingenuity.com'. To choose a different server, enter the server address of desired server. IPA Server address can also be permanently configured from **Tools** → **Options** → **Miscellaneous** → **Pathway Analysis (IPA)**
- **Project Folder:** The name of the project in IPA this pathway should be stored under. By default, it will use the name of the **GeneSpring GX** project.
- **Use both Direct and Indirect relationships:** If selection is 'yes', IPA will build networks using both direct and indirect molecular interactions between genes. If selection is 'no', IPA will build networks using only direct interactions between genes.
- **Knowledge Base content:** This indicates what database will be searched for information to build the network. An empty string indicates all available Knowledge Bases will be searched and information from all sources will be used to perform analysis.
- **Include My Pathways in Enrichment Score:** If selection is 'yes', all pathways saved under My Pathways will be included in the scoring process.

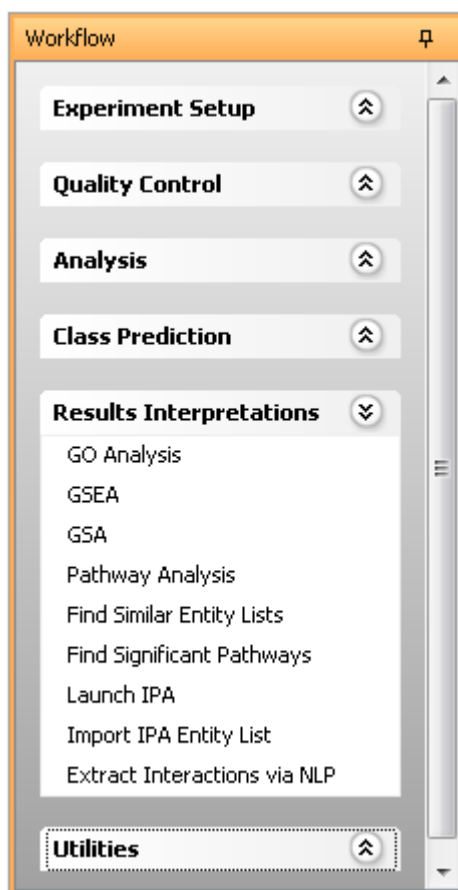


Figure 29.25: Launch IPA

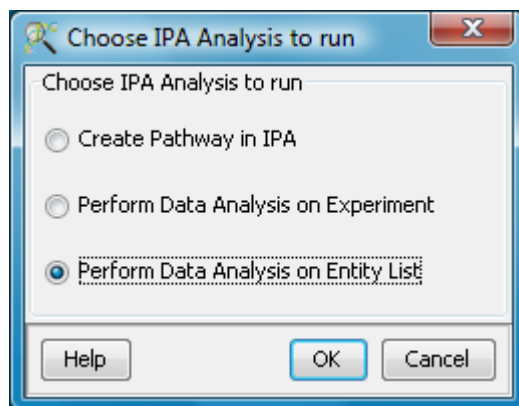


Figure 29.26: Data Analysis on Entity List

- **Review Settings and ID Mapping before Running Analysis:** If selection is 'no', data analysis will be automatically performed using the settings defined in this option. If selection is 'yes', the Create Analysis window below will be displayed for users to review and modify settings before running analysis.
- **Gene Identifier Column:** This indicates the type of gene identifier that will be used to map genes in the Entity List to genes in the Ingenuity Pathways Knowledge Base (IPKB). The type of identifier selected, determines which annotation column in the **GeneSpring GX** technology the identifiers are retrieved from. The list of supported identifiers is given below.

These identifiers will then be used to map genes in the list to genes in IPKB. Only identifiers that can be matched to genes in IPKB will be used to build a new pathway in IPA.

- Entrez Gene ID
- Locus Link ID
- Affymetrix Probeset ID
- UniGene ID
- GenBank Accession
- Swissprot
- Agilent Probe ID
- RefSeq Protein ID

Press **OK** to start the analysis in IPA. Your default browser will start up and connect to the IPA server.

- **Associated Value Column:** If the selected entity list has one or more associated value columns, this drop down contains the names of the columns and one column can be selected.
- **Associated Value Type:** This sets the type of value that is represented by the associated value. For the fold change selection, **GeneSpring GX** will convert the **GeneSpring GX** notation (2, up and 4, down etc.) to the IPA notation (2 and -4). The **GeneSpring GX** notation is stored in 2 columns but will be converted to one number in accordance with the IPA notation. All other settings do not result in any conversion and it is up to the user to choose the correct Value Type.

IPA analysis

The connection to IPA is performed through the IPA HTTP API and on some systems this may result in the session being blocked due to the security settings as shown below. If this happens, simply click on the yellow bar and select 'Allow Blocked Content'. There is no danger in allowing this access. See figure [29.28](#)

IPA is using Java Webstart and a dialog appears indicating JAVA is starting up. See figure [29.29](#).

The login dialog will appear. Enter your login credentials here. See figure [29.30](#)

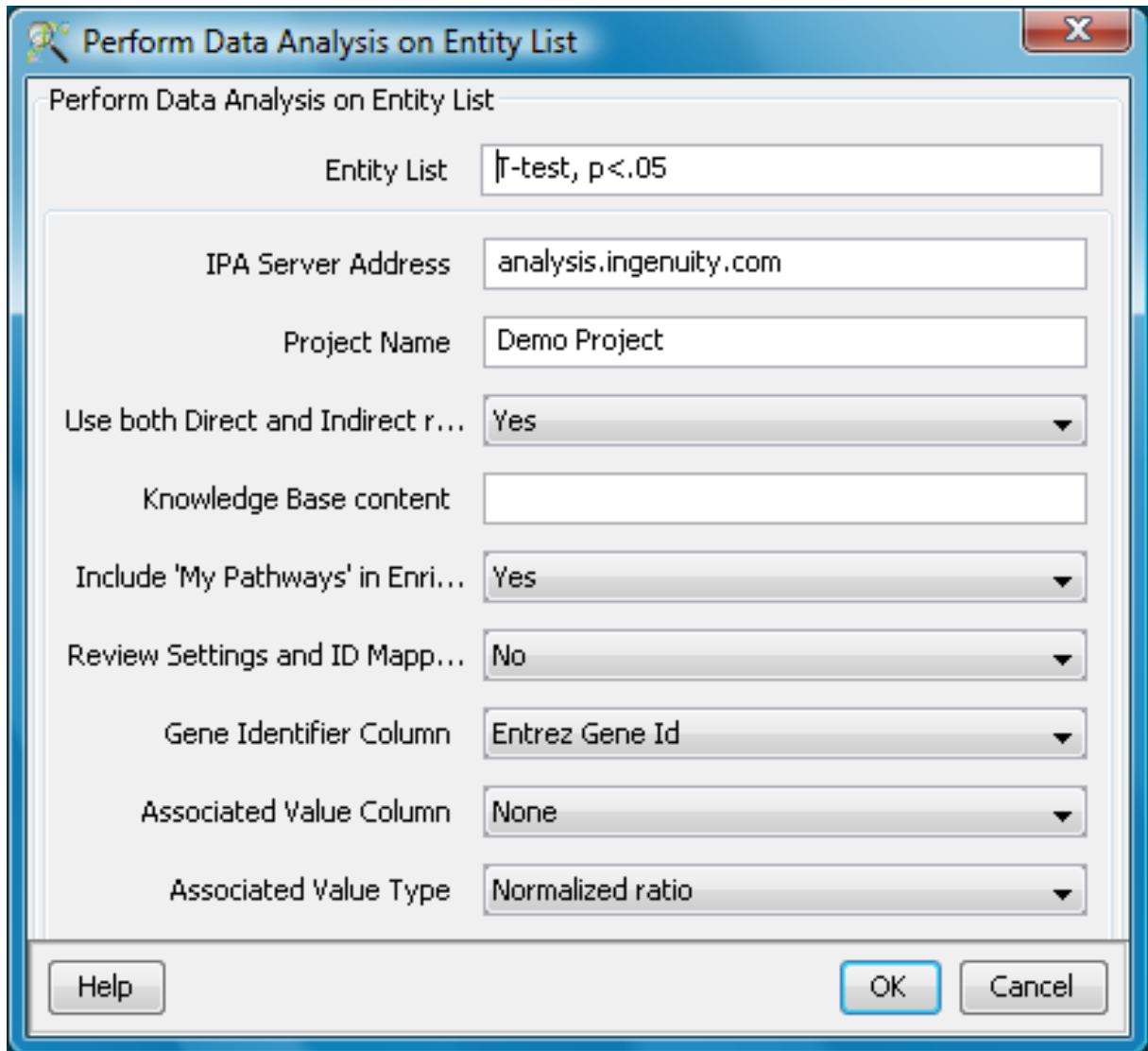


Figure 29.27: Perform Data Analysis on Entity List

If 'Review settings and ID mapping before running an analysis' was selected, the following screen [29.31](#) in IPA will appear, if not, the main window will show.

Review the analysis settings and press 'Run Analysis' to proceed. Review the analysis settings and press 'Run Analysis' to proceed. See figure [29.32](#)

The analysis is performed and the Analysis will have a little clock in the icon while it is running. When the analysis is finished, the clock will disappear the analysis is ready for review.

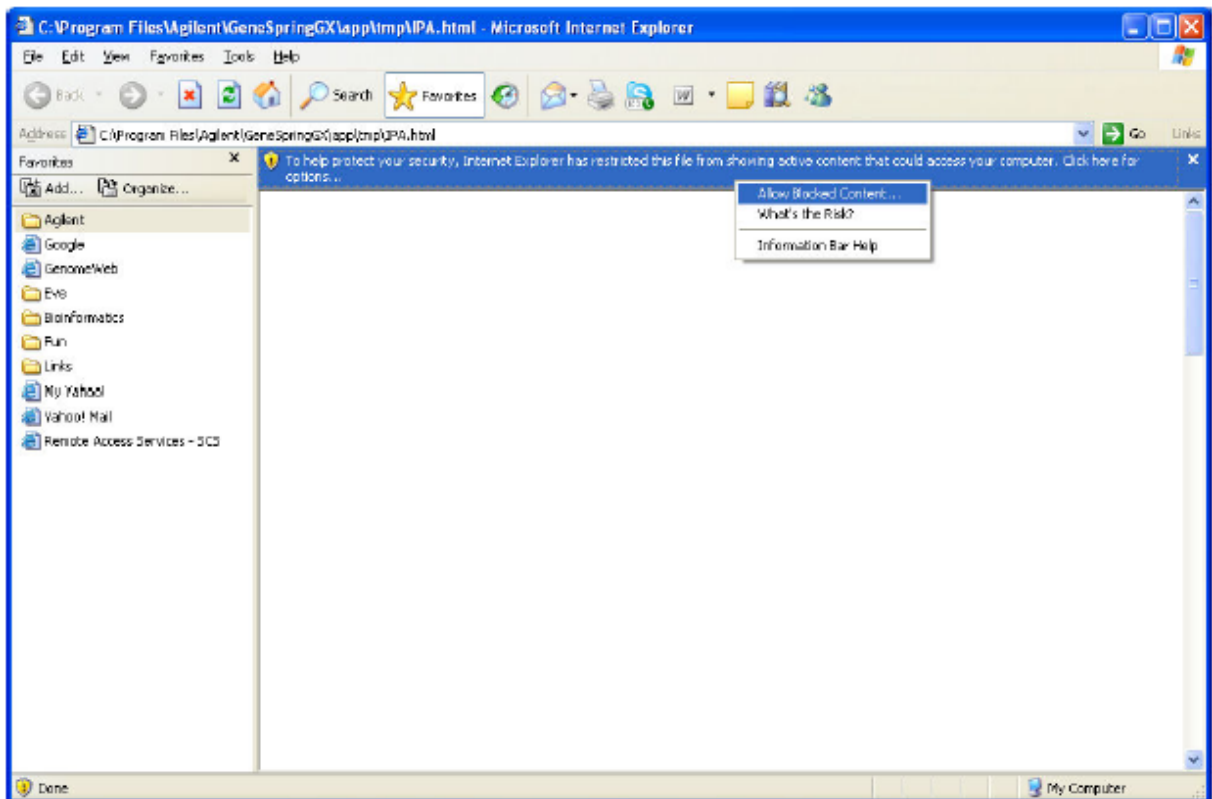


Figure 29.28: IPA Pathway Creation



Figure 29.29: Java Startup

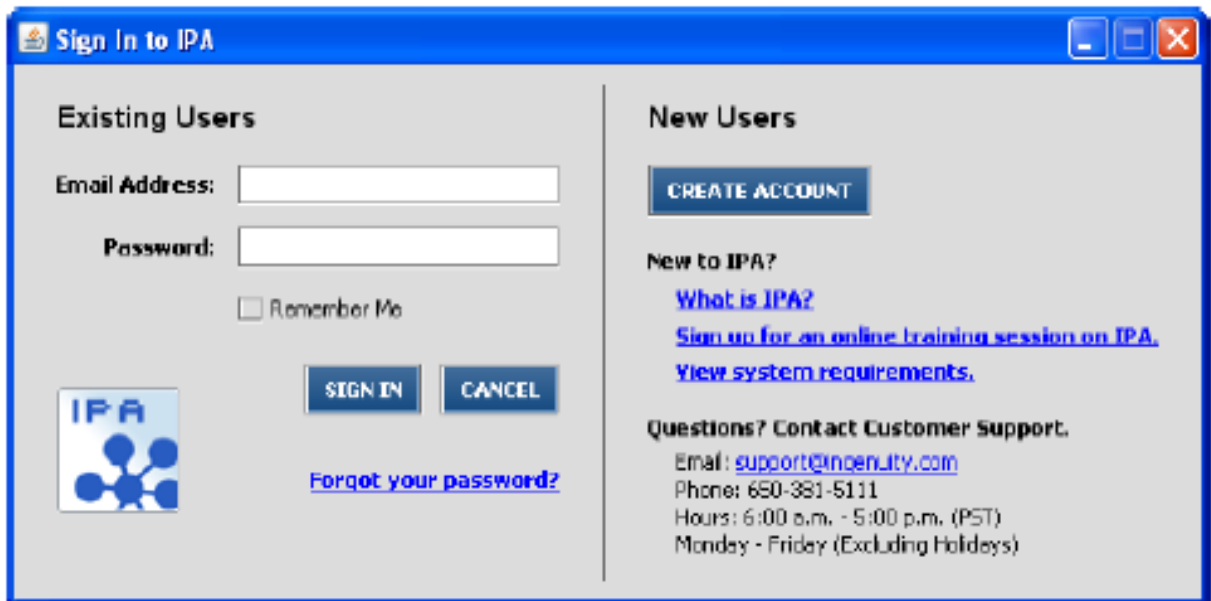


Figure 29.30: IPA Login Dialog

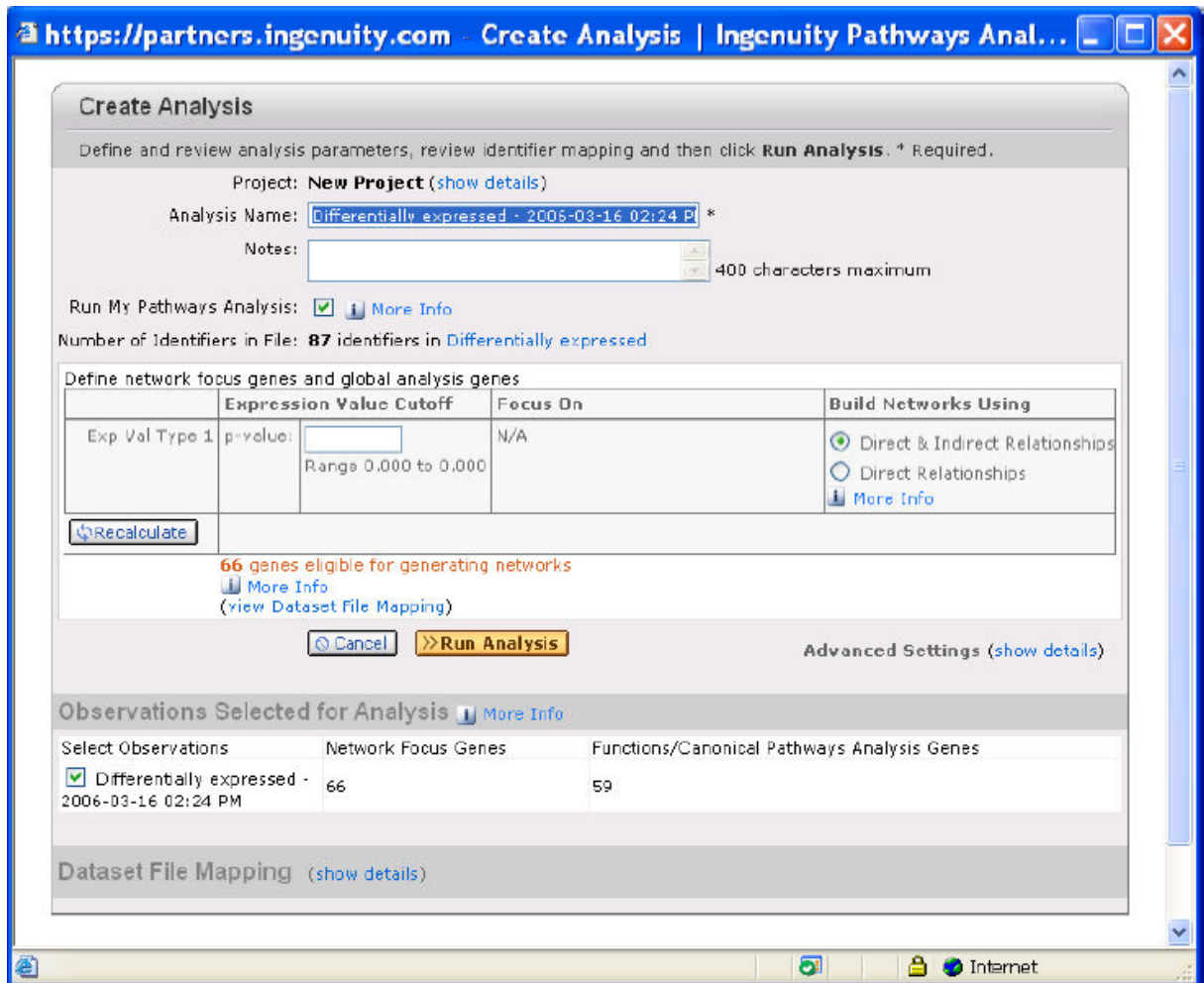


Figure 29.31: Create Analysis

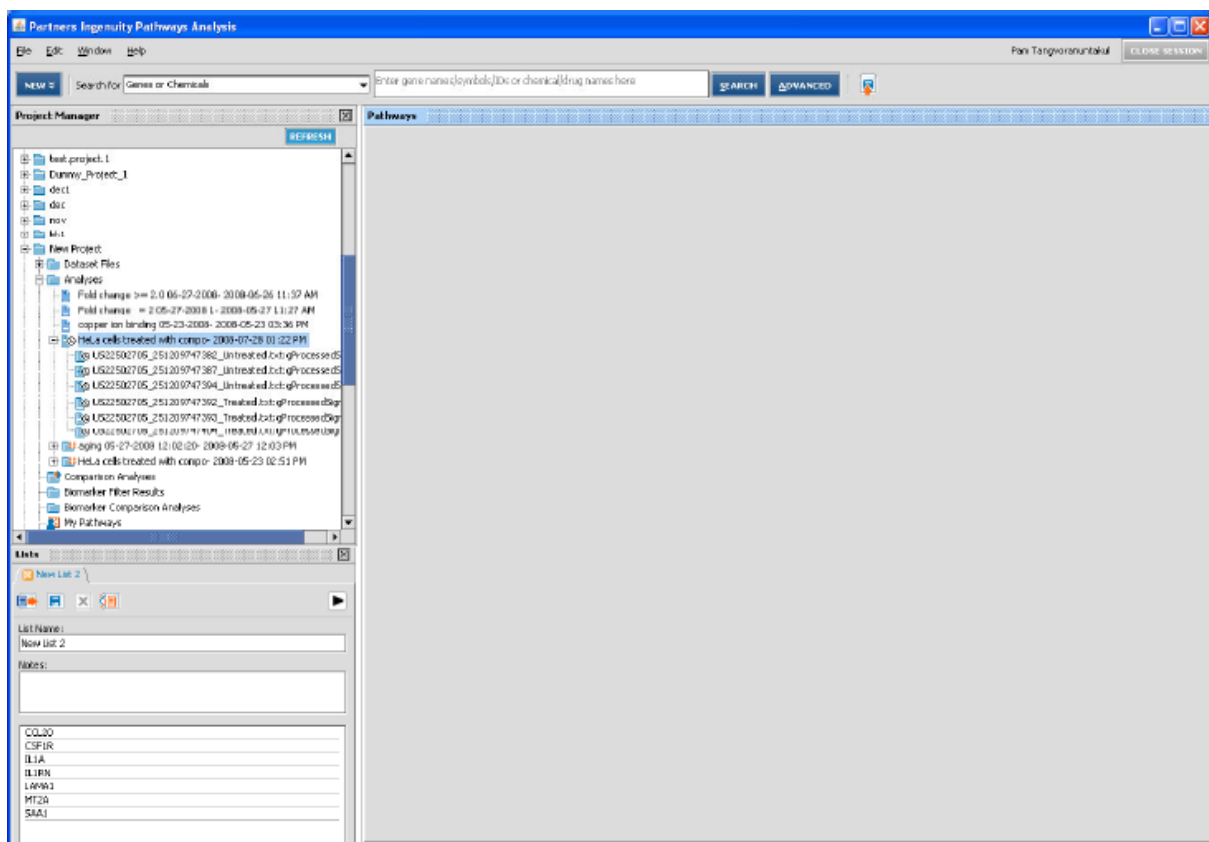


Figure 29.32: Analysis Settings

Chapter 30

GeneSpring GX Workgroup Client

The **GeneSpring GX** is available in two modes, namely, Desktop mode and Workgroup mode. The Desktop mode allows the user to perform the analysis including creation of objects locally on the machine whereas in the Workgroup mode, analysis objects can be shared among the users and groups. Objects can also be created locally as well as remote execution for some operations is possible. The Workgroup mode needs Workgroup Server to be installed in Enterprise along with client licenses provided with it.

30.1 Users and Groups

Consider a simple scenario of an organization where one group of people (say Core-Facility) conducts microarray experiments and another group (say Research) does the analysis. The Core-Facility will perform microarray experiments and store samples on the **GeneSpring Workgroup Server** . The Research group creates Projects and Experiments using the samples provided by Core-Facility and performs analysis. The results in the form of Entity lists, Prediction models, Pathways, Experiments or even the whole Project can be shared with other members of the Research group for further analysis. This introduces the notion of a user and group. Therefore, to share objects among the members of different teams, users and respective groups need to be created on **GeneSpring Workgroup Server** .

The **GeneSpring Workgroup Server** has a notion of user accounts to maintain privacy of data and a notion of groups and permissions to facilitate collaboration with fellow researchers. Access to the **GeneSpring Workgroup Server** from the **GeneSpring Manager** or the **GeneSpring GX Client** is restricted using the password protected user account. The administrator needs to create these user accounts and groups. The **GeneSpring Workgroup Server** can have a number of users, and each user can be part of zero, one or more groups. Each group can have a group administrator who acts as an administrator of the group. All the objects stored on the server can be shared among individual users and groups using the client.

30.1.1 Login

After activating **GeneSpring GX** client , the tool launches and opens up the login dialog.

The following connection parameters to the **GeneSpring Workgroup** Server needs to be specified in the login dialog:

- Host - The hostname or IP address of the machine where **GeneSpring Workgroup** Server is running.
- Port - The port number at which **GeneSpring Workgroup** Server is running. The default port is 8080.
- Login - The user account which the administrator has created for you to connect to the server.
- Password - The password for your user account used for authentication.
- Use SSL - Select this option if you want the communication between client and server to be secure. The port number is 8443.
- Proxy settings - If the **GeneSpring GX** and **GeneSpring Workgroup** Server are on different sides of a proxy server, then enter the proxy settings in the *Proxy* tab of the login dialog. To specify proxy settings check the *Use Proxy* checkbox and enter the hostname or IP address of the proxy server and the port number on which it runs. If you connect to proxy server using a username and password enter it here or leave it blank.

All the connection parameters except the password will be remembered the next time you launch the **GeneSpring GX** Client.

30.2 Operations on GeneSpring Objects

Objects - When a user creates Projects, Experiments, Samples, all of these and other analysis objects are stored on the **GeneSpring Workgroup** Server and can be shared among different users and groups as per the permissions granted by the owner. Refer section on [objects](#).

Independent Objects - are the ones which can be shared independently such as Projects, Experiments, Samples and Entity Lists, Pathways, Prediction models and Scripts. Dependent objects are the ones that can be shared through their parents only like Condition Trees, Row Trees, Combined trees, Interpretations and Classification.

30.2.1 Object ownership

What is ownership: Every object has a creator and an owner associated with it. The creator of an object is the user who created it and can never be changed.

Who can change ownership: The owner of an object can be changed from one user to another several times in the life span of the object. At any instance though, there will be one user who owns the object. Objects cannot be owned by groups. When the object is first created, the creator and the owner are both the same user. The group administrator can change the ownership of any object owned by a member of his group. The administrator can change the ownership of any object in the system.

How can ownership be changed: The ownership of an object can be changed via *Change Owner* option in the right click menu of the object. The *Change Owner* dialog shows the members of the group in the drop down menu. Select the user and click OK to change the ownership of the object.

The *Change Owner* option is available only for independent objects. While changing ownership one can choose to propagate the ownership change hierarchically to all the children of the object as well. A Project's ownership can be changed from *Project* menu.

Delete object: Objects can be deleted via *Delete Object* option in the right click menu of the object or from *Delete Object* icon given in the toolbar menu of the search wizard. Only the owner, group administrator or the administrator has the authority to delete objects.

30.2.2 Object permissions

What are permissions: Objects can be given two kinds of permissions for sharing.

1. Read

Objects with read permission to a group are accessible to the members of the group. The members can search for these objects, inspect these objects, open them in case of Projects and Experiments, or use them in case of Models, Scripts, Pathways, Samples and Entity lists. However, no modification can be done to the objects.

2. Write

Objects with write permission for a group can be accessed and also modified by members of the group. Modifications include changing object attributes, adding children (Experiments to Projects or Entity Lists to Experiments), edit attachments to Samples, changing contents of Scripts etc.

If an object has neither read nor write permissions for a group, then the members of the group cannot access the object. Permissions can also be explicitly set for an individual user. If a user has explicit read or write permission for an object, then that user can access or modify that object respectively.

Who can change permissions: Only the owner of the object or administrator or group administrator has the authority to change permissions.

How can permissions be changed: The permissions can be changed from the *Share object* option in the right click menu of the objects or the from *Change Permissions* icon in the search wizard.

To share an object, right-click on the object and select *Share*. This brings up the *Permissions* dialog. See Figure 30.1. The upper panel shows the object selected and the lower panel shows the current state of permissions on the chosen object. The *Group* tab shows the groups to which the user belongs and the explicit permissions to each group. The *User* tab shows all the users in the groups that the owner belongs and the explicit permissions given to each user. Multiple objects can be shared by selecting multiple objects and click *Share Objects* from the right click menu or *Change Permissions* icon in the toolbar menu of the search wizard. In this case the upper panel shows all the independent objects chosen. The lower panel shows the permissions on these objects for each user/group if all these permissions match.

Permissions to Folders, Projects, Experiments, and analysis objects can be propagated to all its descendents in the hierarchy.

Note that the interface shows the explicit permissions granted by the owner of the object. If a group has explicit permissions to read/write on an object, all users in that group will also have read/write permissions implicitly.

Object permissions to all users (even outside the groups that the user belongs) in the system can be given by sharing objects with *Everyone* group. Depending on the policies of the organization, **GeneSpring Workgroup** Server administrator can allow or disallow sharing objects to Everyone.

30.2.3 Conflicts with permissions

There can be two kinds of conflicts regarding permissions

User-Group: If a group is given write permission for an object by its owner but one of its members is explicitly given read permission then the permission granted by the owner explicitly for that particular user will take precedence while for other members write permission will be retained.

Group-Group: If a user belongs to two groups and an object has read permission to one group and write permission to another then the most liberal permission i.e. write permission will take precedence for user A on that object.

30.2.4 Propagating permissions

- When an object is shared using right click option or search menu, share dialog opens up displaying the users and groups. Clicking OK on the *Permissions* dialog prompts a window asking whether to propagate the permissions of the parent to its children or not. If the permissions of the parent are not propagated to its children then they retain their original permissions.

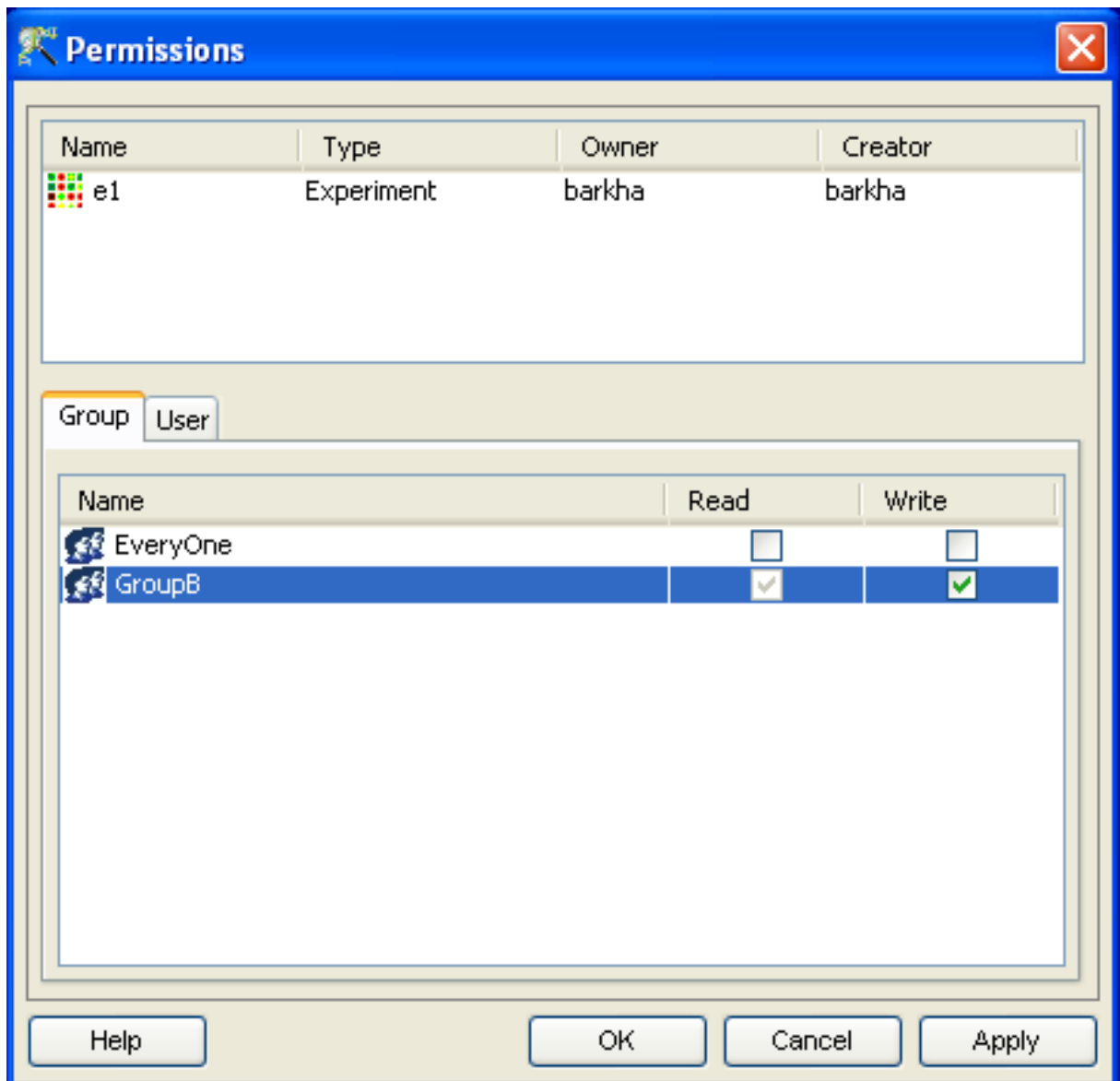


Figure 30.1: Permission Dialog

- When permissions are propagated to children, it is the change in permission that gets propagated. Consider a Project with Group A:read , that contains an Experiment with Group A:read; Group B:write with propagation. If the Project is changed to Group A:write with propagation, the experiment permission will change to Group A:write; Group B:write. Thus the change in permissions (Group A:read \rightarrow Group B:write) is what gets propagated to all the children. Other permissions if any on the children are retained.
- While propagating permissions, it may be that there are some children on which permissions cannot be altered by this user. Consider a Project owned by user A containing two Experiments, one owned by user A and another by user B. When user A changes permissions on the Project and propagate permissions, then the propagation applies only to the Experiment owned by him. The permissions on the Experiment owned by user B are not changed since user A does not have privileges to do so.

30.2.5 Inheriting Permissions

- When existing objects are added to a parent, then the permissions on the object do not change. Thus if an existing Experiment is added to a project or if an Entity List is added to an Experiment, the original permissions of Experiment on Entity List are retained.
- When a new object is created and added to a parent, the object inherits a projection of permissions from the parent. Consider a Project owned by user A, with Group A:write; Group B:read. If the user A creates a new Experiment and adds it to the project simultaneously, then the Experiment will completely inherit the permissions of the Project namely Group A:write; Group B:read. The same holds true if another user B who belongs to both groups Group A and Group B creates and adds a new Experiment. If a user C who belongs to Group B creates and adds a new Experiment, then the Experiment will carry only a projection of permissions from the Project namely Group B:read. The Group A:write permission is not inherited since user C does not belong to Group A. In addition, the owner of the Project user A, gets write permissions on the Experiment.

30.3 Remote Execution

In **GeneSpring Workgroup Server** , the user can execute certain resource intensive tasks remotely on the **Compute Server**. This allows the user to perform other analysis on his machine. Powerful machines are usually designated for Compute Server(s) and configured to enable execution of several such tasks simultaneously.

30.3.1 Task Manager

The **Task Manager** acts as an interface between the **Compute Server** and the client by providing the status and other details of the remotely executed tasks. It can be accessed via *Tools \rightarrow Task Manager*. The Task Manager displays a table with the following columns: Position in Queue, Name, Status, Owner,

Submission Time, Scheduled Time, Start Time, End Time, Compute Server IP, Application Context and TaskID. If the task execution has started, the status message shows **Executing**. In case the task is scheduled for a later time, the status message shows **Scheduled**. And in case of **Compute Server** not being available to perform the task at the specified time, the status shows **Queued**. The **Position in Queue** for all **Scheduled** tasks is 0 and the **Queued** tasks are allotted numbers starting from 1. As the tasks start getting executed, the position in queue moves up for the remaining **Queued** tasks. Once the remote task has finished execution, the status will be changed to **Successful** or **Failed**. Typically all tasks go through the following order: **Scheduled** → **Queued** → **Executing** → **Successful**. The Task ID is a unique number allotted to each task and is used in the internal machinations of the **Compute Server**.

A normal user can only see his/her remotely executed tasks on the **Task Manager** whereas the administrator can see all the tasks of all the users on his/her **Task Manager**.

GeneSpring Workgroup Server allows several operations to be performed on the tasks in the **Task Manager** depending on their status such as

- Scheduled tasks can be rescheduled to a different date and time or can be suspended or deleted to prevent them from getting **Queued**. Suspension of tasks is temporary in nature and can be resumed whereas deletion is permanent.
- Queued tasks can be suspended (temporarily) or terminated (permanently) to prevent them from getting executed.
- Suspended tasks can be resumed. If the scheduled time of the task has passed, then the task will be **Queued**, else it will **Scheduled**
- Executing tasks can be terminated. This will stop the execution of the task on the **Compute Server** and change the status to **Deleted**
- Successful/Failed/Deleted tasks can be cleared to remove them from the **Task Manager** and move them to the **Task History**.

The table in the **Task Manager** can be updated by clicking on the **Refresh** button. The logs can be viewed for **Successful** and **Failed** tasks and help the user in understanding the reason behind the successful/unsuccessful execution. The **View Status** tab allows the user to see the memory usage of the **Executing** task on the **Compute Server**. See the Figure [30.2](#).

30.3.2 Remotely Executable Operations

- **Experiment Creation:** The experiment creation task can be executed remotely by checking the **Execute remotely** option in the last step of the experiment creation wizard in Advanced Analysis workflow mode. If the experiment is being created from raw files, they are first uploaded from the local machine as samples and then the experiment creation task is scheduled remotely.

When the task is successful, the result of the task(a new experiment) can be viewed by refreshing the project, if it is already open.

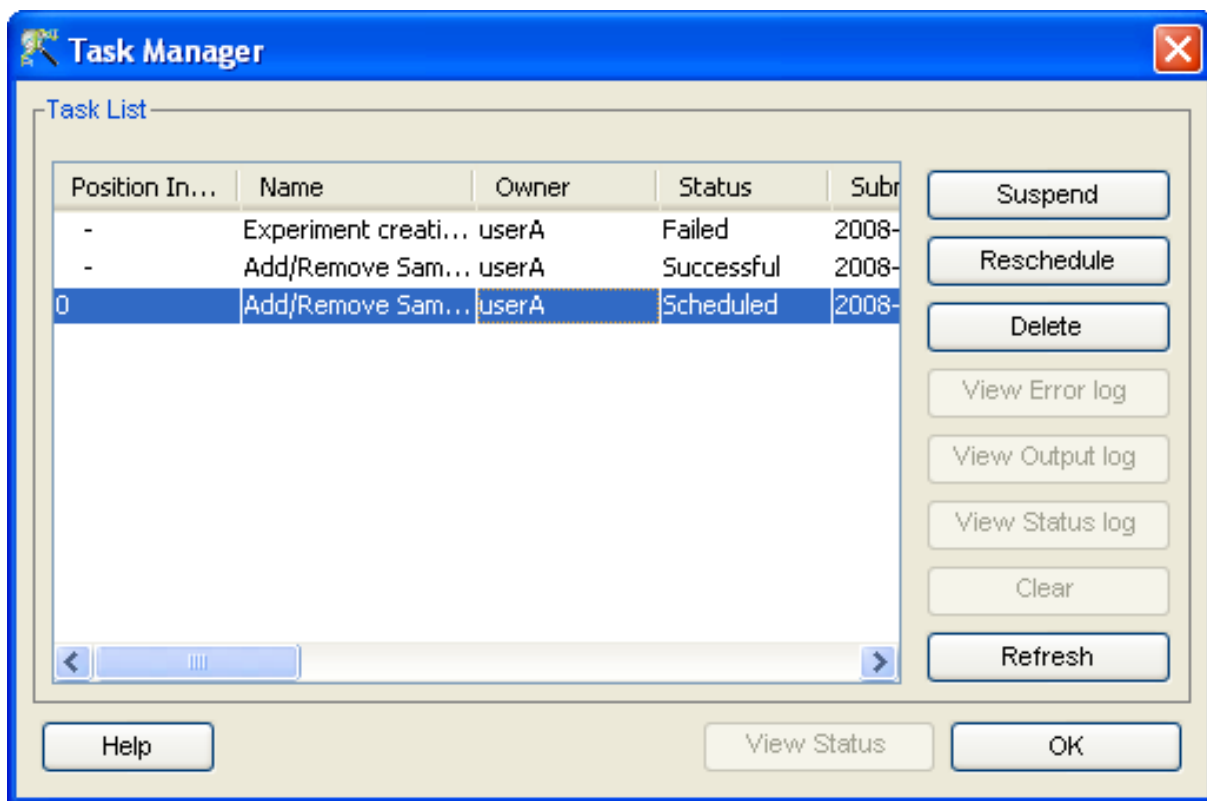


Figure 30.2: Task Manager

- **Add/Remove Samples:** The Add/Remove Samples option can be accessed via *Quality Control* → *Quality Control on Samples* link in the workflow browser in the Advanced Workflow and in the Guided Workflow, it can be accessed in the Quality Control step(3rd step).

After clicking on the Add/Remove Samples button and selecting samples for addition/removal, check the **Execute remotely** option to execute this task remotely on the **Compute Server**.

When the task is successful, the result of the task can be viewed by refreshing the experiment, if it is already open.

- **Clustering:** Clustering can be performed in the Advanced mode via *Analysis* → *Clustering* from the workflow browser. After selecting the various options for performing Clustering, the second step has the option to perform remote execution.

When the task is successful, the result of the task can be viewed by refreshing the experiment, if it is already open.

- **SNP Tagging:** SNP Tagging can be performed remotely from the first page of the SNP Tagging Wizard. You can configure the SNP Tagging parameters (refer to [SNP Tagging](#) section) from this page, and run the analysis in the Workgroup Server.

1. Select the **Execute remotely** check box. This will provide two options:
 - Select the **Schedule Now** radio button to start the analysis immediately. If there is a pre-existing queue of analysis, then the process will start immediately after the last analysis in the queue.

- Select the **Schedule task** at radio button, and then select a time and date from the drop-down option to run the analysis at a desired time.
- 2. Click Next. This will open an "Information" message confirming the task schedule. You can check the status (refer to [Task Manager](#) section) of the task from em Tools → *Task Manager*.
- 3. Once the task is finished, refresh the project. You will be able to see a new entity list of **Tagged SNPs** under the **All Entities** list.

30.3.3 Interpreting Task Logs

:

- While executing any of the remote tasks if the Compute Server runs out of memory, then the task does not get executed and the task status in the **Task Manager** shows **Failed** and the **View Status log** shows that the Compute Server has run out of memory. This can be modified by the administrator of the Compute Server.
- The project gets deleted before or during the remote experiment creation. In that case, the task status in the **Task Manager** shows successful and the **View Status log** displays a message that the experiment creation was successful but the experiment was not added to the project. However the created experiment can be viewed via *Search* → *Experiments* and the particular experiment can be added to the desired project(which has to be open at that time) using the icon for the same.
- The permission to the shared project is revoked by the owner while another user has scheduled a remote experiment creation on the same project. Then the **Task Manager** shows that the task was successful and the **View Status log** shows that the task was successful but the experiment is not added to the project. But the experiment can be added to another project by doing *Search* → *Experiments* from the tool bar and then adding the particular experiment using the icon for the same.
- The permission to the shared samples is revoked or the owner deletes the samples while another user is creating an experiment remotely with them. Then the task status in the **Task Manager** shows **Failed** and the **View Status log** shows that there was an error in reading the sample.
- The experiment gets deleted when the task 'Add/Remove Samples' is getting executed remotely. In that case, the task status shows **Failed** and the **View Status log** shows that the operation failed unexpectedly.
- The task 'Add/Remove Samples' is scheduled for remote execution on a shared experiment for which the user has only **Read** permission. Then the task status shows **Failed** and the **View Status log** shows that there was an error in reading the experiment.
- When performing Clustering remotely with 2 samples using K Means and with the following parameters- Cluster on Conditions and Number of Clusters is greater than 2 then the task status shows **Failed** and the **View Status log** shows that the number of clusters are same or more than the number of columns.



- The experiment gets deleted after scheduling Clustering remotely. Then the Task Status shows **Failed** and the **View Status log** shows that the particular entity list on which Clustering is to be performed could not be found and might be deleted.
- The task 'Clustering' is scheduled for remote execution on a shared experiment for which the user has only **Read** permission. Then the task status shows **Failed** and the **View Status log** shows that there was an error in adding classification to the experiment



Chapter 31

Writing Scripts in GeneSpring GX

GeneSpring GX offers users the ability to write their own scripts, execute these within the current context and save them. The scripting utility allows users to combine operations in **GeneSpring GX** with a more general Jython programming framework to yield automated scripts. Using these scripts, one can run transformation operations on data, automatically pull up views of data, and even run algorithms repeatedly, each time with slightly different parameters.

31.1 The Script Editor

Scripts can be typed, or copied and pasted, into the Script Editor Window. To run a script, go to *Tools* → *Script Editor*. This opens up the Script editor window. Write your python script into this window and click on the Run  icon to execute the script. Alternatively, if you have a script in a file, open the file in any editor, copy and paste the script into the script editor and click on Run  icon to execute the script. Errors in the execution of this script, if any, will be recorded in the log window, which acts as the default standard output location.

Once the script is in the window, it can be saved using the Save  icon . Clicking on the Save icon opens up a Script inspector where the name of the script can be given. It also contains the creation date, modification date and the owner. Click OK to save the script. The scripts are saved into the Workgroup database by default and can be retrieved later using the Open  icon icon.

This chapter provides API documentation for some of the basic operations needed to access, manipulate and save data in **GeneSpring GX**, run algorithms, generate customized views and save these results. We also have a section detailing how you can interface these options with R packages available outside the tool.

Please note that the Script Editor in **GeneSpring GX** expects Jython (Python with Java class im-

port capabilities) scripts. Therefore, to write meaningful scripts using the functions described in the documentation below, you will need some knowledge of the Python programming language. See <http://www.python.org/doc/tut/tut.html> for a basic Python tutorial. Knowledge of Java will help you write more powerful scripts. See <http://java.sun.com/j2se/1.5.0/docs/api/> for the Java API.

31.2 Hierarchy of data organization in GeneSpring GX

In order to access data within **GeneSpring GX**, it is essential to understand the data organization hierarchy. Refer to the section on [Data Organization](#). At any point in time, **GeneSpring GX** can have only ONE active project, which can contain one or more experiments, out of which only ONE will be the active experiment. Each experiment will have its own Samples, Interpretations and Analysis.

31.2.1 Accessing Projects, Experiments and their Constituent Elements

Usage	Notes
<pre>#Get the current active project p = script.marray.project.getActiveProject()</pre>	The active project (object) gets stored in the “proj” variable
<pre>#Get the currently active experiment e = script.marray.project.getActiveExperiment()</pre>	The active experiment (object) gets stored in the “exp” variable
<pre>#Get the list of Samples in this experiment samples = expt.getSamples()</pre>	“samples” variable will contain a list of the samples using which this experiment was created. The expt variable should contain an experiment object as shown in (2) above.
<pre>#Get the parameters for a sample params = sample.getParameters()</pre>	“params” will now contain a java LinkedHashMap which gives details of the sample such as the names of the experiments it is grouping information (which class it belongs to in each interpretation in that experiment).
<pre>#Get the default interpretaion for this experiment interpretation = expt.getDefaultInterpretation()</pre>	The expt variable should contain an experiment object as shown in (2) above
<pre>#Get the active interpretation for this experiment interpretation = expt.getActiveInterpretation()</pre>	The expt variable should contain an experiment object as shown in (2) above

Continued on Next Page...

Table 31.1 – Continued

Usage	Notes
<pre>#Get the active technology for this experiment techId = expt.getTechnology() t = script.marray.project.createTechnology(techId)</pre>	The expt variable should contain an experiment object as shown in (2) above. The “technology” variable will now contain the technology object for the experiment “expt”.
<pre>#Get all entity lists in the experiment entityLists = expt.getAllEntityLists()</pre>	The expt variable should contain an experiment object as shown in (2) above
<pre>#Get the current active entity list in the experiment entityList = proj.getActiveEntityList()</pre>	The proj variable should contain a project object as shown in (1) above
<pre>#Get the “All Entities” entity list entityLists = expt.getAllEntityLists()</pre>	The expt variable should contain an experiment object as shown in (2) above
<pre>#Get the Experiment Grouping exptGrouping = expt.getExperimentGrouping()</pre>	The expt variable should contain an experiment object as shown in (2) above
<pre>#Get the conditions map from the experiment grouping conditionsMap = exptGrouping.getConditionsMap()</pre>	conditionsMap is a java Map object. You can iterate through it as follows: for key in conditionsMap.keySet(): print key, “:”, conditionsMap.get(key)

Table 31.1: Accessing Projects and Experiments

* Please note that there are multiple correct ways to access a function belonging to a module in Python, such as:

1. Including the module in the script:

```
# include the getActiveProject function from the module
script.marray.project in this script
from script.marray.project import getActiveProject
```

```
proj = getActiveProject()
```

Or:

```
# include ALL the functions from the module
script.marray.project in this script
from script.marray.project import *
```

```
proj = getActiveProject()
```

2. Referring to the function in the context of its module:

```
proj = script.marray.project.getActiveProject()
```

31.2.2 Accessing the Experiment Dataset

The data seen in the Spreadsheet view when the “All Entities” Entity list is active in the Project Navigator) shows the master dataset for that experiment, also known as the normalized dataset. This normalized dataset can be accessed as follows:

```
e = script.marray.project.getActiveExperiment()
analysis = expt.getAnalysis()
n = script.marray.project.analysis.getNormalisedDatasetNode(analysis)
ds = ndn.getDataset()
```

The variable “dataset” now contains the normalized dataset for the entire experiment. (You can also replace `getNormalisedDatasetNode` by `getRawDatasetNode` to get the raw dataset.)The dataset contains columns and rows. A number of operations can be carried on a dataset object `ds`, as shown below:

Usage	Notes
<pre>#Get the number of columns in the dataset numCols = ds.getColumnCount()</pre>	The variable <code>numCols</code> will contain number of columns (both data and annotation columns) in the dataset <code>ds</code> .
<pre>#Get the number of rows in the dataset numRows = ds.getRowCount()</pre>	The variable <code>numRows</code> will contain number of columns in the dataset <code>ds</code> .
<pre>#Get the ith column in the dataset column = ds.getColumn(i) pythonCol = script.coercion.to_py(column)</pre>	Here the dataset variable must contain a valid dataset. <code>i</code> must be an integer value between 0 and <code>n-1</code> where <code>n</code> is the number of columns. This will get the data from the column into a python list.

Continued on Next Page...

Table 31.2 – Continued

Usage	Notes
<pre>#Get the contents of the ith col, jth row in the dataset ds data = ds.getColumn(i).get(j)</pre>	Gets the data from the ith column, jth row.
<pre>#Create new integer column data = [1,2,5,6] # data can be a python list name = intColumn ic = script.dataset.createIntColumn(name, data)</pre>	Creates a Integer column with the specified name having the given data as values. Here, name is a string, data is an python list containing integer values.
<pre>#Create new float column fc = script.dataset.createFloatColumn(name, data)</pre>	Creates a Float column with the specified name having the given data as values. Here, name is a string, data is a python list containing float values.
<pre>#Create new string column sc = script.dataset.createStringColumn(name, data)</pre>	Creates a String column with the specified name having the given data as values. Here, name is a string, data is a python list containing string values.
<pre>#Create a new dataset cols = [] cols.append(intCol) cols.append(floatCol) cols.append(stringCol) d = script.dataset.createDataset("NewDataset", cols)</pre>	<p>(1) Create an array of columns (2) Populate it; append previously created columns (see above) to the array. Alternatively, you may use a python list:</p> <pre>cols = [iCol, fCol, sCol]</pre> <p>(3) Create a new dataset consisting of those columns.</p>
<pre>#Add a new column to the dataset ds ds.addColumn(col)</pre>	Adds the “col” column to the dataset ds
<pre>#Transpose the dataset ds from com.strandgenomics.cube.dataset import DatasetUtil transDs = DatasetUtil.transpose(ds)</pre>	transDs will contain the transpose of ds
<pre>#Get all the data from a column into a python float array (float []) from com.strandgenomics.cube.dataset import DatasetUtil from com.strandgenomics.cube.framework.data import ArrayUtil strandFloatArray = DatasetUtil. getColumnData(ds.getColumn(n)) floatArray = ArrayUtil.getContents(strandFloatArray)</pre>	The variable floatArray will contain all the data in the nth column in the dataset ds. n must be an integer.

Continued on Next Page...

Table 31.2 – Continued

Usage	Notes
<pre>#Get all the data from a column into a python integer array (int[]) from com.strandgenomics.cube.dataset import DatasetUtil intArray = DatasetUtil.getIntegerArray(ds.getColumn(n))</pre>	<p>The variable intArray will contain all the data in the nth column in the dataset ds.</p>
<pre>#Get all the data from a column into a python string array from com.strandgenomics.cube.dataset import DatasetUtil intArray = DatasetUtil.getStringArray(ds.getColumn(n))</pre>	<p>The variable stringArray will contain all the data in the nth column in the dataset ds.</p>

Table 31.2: Accessing Experiment Dataset

31.2.3 Some More Useful Functions

Usage	Notes
<pre>#Create Entitylist of non missing signal values #for given experimentcreate Entitylist of non #missing signal values for given experiment script.marray.project. createNonMissingSignalsEntityList(expt)</pre>	<p>A new entity list, “Entities without any missing signal values” is added to the experiment “expt” and will be seen in the Project Navigator.</p>

Continued on Next Page...

Table 31.3 – Continued

Usage	Notes
<pre>#Create New Entity List entitylist = script.marray.project.createNewEntityList (name, notes, entityListType, entityListTechnology, ds)</pre>	<p>A new entity list is created and stored in the variable “entitylist”. “name” and “notes” are string variables of your choice. entityListType and entityListTechnology can be obtained, if you choose, from the current experiment “exp” as follows:</p> <pre>entityListType = exp. getExperimentType() entityListTechnology = experiment. getTechnology()</pre> <p>‘ds’ is the dataset you want to save in your entity list and can be created as shown previously. ds should have the first column as the identifier column with identifiers from the technology.</p>
<pre>#Add an entity list to the experiment exp parentList = expt.getAllEntitiesEntityList() parentList.addChild(entityList)</pre>	<p>Here, instead of the AllEntities List, you could choose to have some other list as “parentList”. See <code>expt.getAllEntityLists()</code> above.</p>
<pre>#Get the flag dataset expt = script.marray.project. getActiveExperiment() analysis = expt.getAnalysis() flagDs = script.marray.project. analysis.getFlagDatasetNode(analysis). getDataset()</pre>	<p>This gives flagging information about each data point in the dataset</p>
<pre># Get the indices of all signal columns # in the dataset sigColIndices = script.marray.project. analysis.getSignalColumnIndices(ds)</pre>	<p>sigColIndices will contain an array giving column indices of the signal columns in the ds dataset.</p>

Continued on Next Page...

Table 31.3 – Continued

Usage	Notes
<pre># Get the column having a certain mark # (such as GO ID column). # Each column in a dataset will have a particular mark, # which identifies its type i.e. Signal column, # identifier column, annotation column etc. from com.strandgenomics.spring.project.mark import MarkManager marks = MarkManager().getMarks() markType = "" for mark in marks: if mark.toString() == "Gene Ontology accession": markType = mark.getType() break col = script.marray.project.analysis. getColumnOfMarkType (ds, markType)</pre>	<p>1) Get the list of all marks. You could print this to see which mark matches your needs, and use it for the match in (3)</p> <p>(2) Create a variable markType.</p> <p>(3) If the mark equals “Gene Ontology accession”, get the corresponding markType</p> <p>(4) Get the column with the correct markType for Gene Ontology accession, i.e get the Column with GO ID values in the dataset ds. (See table below for some common marks*)</p>
<pre># Get column indices for an interpretation colIndices = script.marray.project. translate.getColumnIndices(interpretation)</pre>	
<pre># Get annotation columns for a # particular technology names = technology.getAnnotationColumnNames() cols = script.marray.project.ui. getAnnotationColumns(technology, names)</pre>	<p>“technology” is a variable containing the technology.</p>
<pre>#Get the default annotations dataset techId = expt.getTechnology() techDataset = script.marray.project.ui. getDefaultAnnotationsDataset(techId)</pre>	<p>techDataset will contain the default annotations for the technology with technology ID techId</p>
<pre>#Show a profile plot for the dataset ds script.view.ProfilePlot(dataset = ds, title = My Profile Plot, columnIndices = colIndices). show()</pre>	<p>ds and colIndices should be appropriately populated, as shown in the examples above. You can also manually create a colIndices array as follows:</p> <pre>colIndices = (1,2,5,8,10)</pre>

Continued on Next Page...

Table 31.3 – Continued

Usage	Notes
<pre>#Show a histogram for the dataset ds script.view.Histogram(dataset = ds, title = 'My Histogram', column = 2).show()</pre>	ds should be appropriately populated.
<pre>#Show a box whisker plot for the dataset ds script.view.BoxWhisker(dataset = ds, title = 'My Box Whisker Plot', columnIndices = colIndices).show()</pre>	ds and colIndices should be appropriately populated, as shown in the examples above. For a list of all available views, please refer to the “script.view” section in the GeneSpring GX Scripting manual.

Table 31.3: Some Useful Functions

31.2.4 Some Common Marks

Mark name	Signifies
Agilent Probe ID, Identifier, Probe Set ID, Transcript Cluster ID, Exon ID, Affymetrix ProbeSet Id	The identifier for that particular technology.
Normalized Signal, Raw Signal, Background Corrected Signal	Marks for columns containing signal values.
Gene Name, Gene Symbol	Marks for name of the gene.
Gene Ontology accession	GO ID column
Entrez Gene Id	Entrez Gene Id column
Ensembl	Ensembl Id column
Unigene Id	Unigene Id column
SwissProt	SwissProt Id column
Flag	Flag column

Table 31.4: Some Common Marks

Note: Marks are case-sensitive. You can get a full list of available marks by doing the following:

```
from com.strandgenomics.spring.project.mark import MarkManager
marks = MarkManager().getMarks()
print marks
```

Now that you can access the dataset of your choice, or create it using the available datasets within your project, it can be presented using a number of pre-created views, such as Scatter Plots, Box Plots, Histograms etc.. Similarly, a number of popularly used algorithms can be directly run on a dataset, and the results can then be accessed. The API for these functionalities are available in the Scripting Manual here.

31.2.5 Creating UI Components

Using the createComponent function, a UI dialog can be presented to a user so he or she can provide inputs at program runtime. The types of components that you can create are shown below. Please note that in order to use any of these, you need to include the following line in your script:

```
from script.omega import createComponent, showSimpleDialog
```

Usage	Notes
<pre>#Drop down box p = createComponent(type="enum", id="name", description="Enumeration", options=["option 1","option 2","option 3"]) props = {'title' : 'Title for the dialog', 'hasBanner' : Boolean(1), 'bannerDescription' : 'You can write the description for this process here.', 'cancelLabel' : 'Close', 'hasOK' : Boolean(1)} result=showSimpleDialog(p, properties=props) print result</pre>	<p>(1) Creates a UI component p of type “enum” (referring to a drop down box) with 3 options (2)Creates a properties map for the dialog that will show up. (3)Shows the dialog to the user (4) prints result obtained</p>
<pre># Check Box p = createComponent(type="boolean", id="name", description="CheckBox")</pre>	

Continued on Next Page...

Table 31.5 – Continued

Usage	Notes
<pre># Radio Button p = createComponent(type="radio", id="name", description="Radio", options=["option 1","option 2","option 3"])</pre>	
<pre>#File chooser p = createComponent(type="file", id="name", description="FileChooser")</pre>	
<pre># Single column chooser p = createComponent(type="column", id="name", description="SingleColumnChooser", dataset=ds)</pre>	Use this to select one column from the dataset ds
<pre># Multiple column chooser p = createComponent(type="columnlist", id="name", description="MultipleColumnChooser", dataset=ds)</pre>	Use this to select multiple columns from the dataset ds
<pre>#Entity List chooser p = createComponent(id="entityList", type="EntityList", description="Entity List")</pre>	
<pre># Text Box p = createComponent(type="text", id="name", description="TextArea", value="Enter text here")</pre>	You can replace “int” by “float” or “string” in the type field.
<pre># Group 2 or more of the previous components groupComponent = createComponent(id="chooser", type="group", description="Choose your input parameters", components=[c1,c2])</pre>	The 2 components c1 and c2, which can be examples of any of the components shown above, are grouped together and shown in the same dialog to the user.

Table 31.5: Creating UI Components

31.2.6 Example Scripts

```
##### Example 1 #####
#
# script to choose column to display an MVA plot
#

from script.view import ScatterPlot

from script.omega import createComponent, showSimpleDialog

exp = script.marray.project.getActiveExperiment()
analysis = exp.getAnalysis()
ndn = script.marray.project.analysis.getNormalisedDatasetNode(analysis)
d = ndn.getDataset()

#
# define a function for opening a dialog
#

def openDialog():
    A = createComponent(type='column', id='column A', dataset=d)
    B = createComponent(type='column', id='column B', dataset=d)
    C = createComponent(type='column', id='color by', dataset=d)

    g = createComponent(type='group', id='MVA Plot',
                        components=[A, B, C])

    result = showSimpleDialog(g)

    if result:
        return result['column A'], result['column B'], result['column C']
    else:
        return None

#
# define a function to show the plot with two columns of the
# active dataset and show the results
#

def showPlot(avg, diff, color):

    plot = script.view.ScatterPlot(title = 'MVA Plot', xaxis=avg,
                                   yaxis=diff)
    plot.colorBy.columnIndex = color
```

```

plot.show()

#
# main
# This will open a dialog, and take inputs
# Compute the average and difference
# Append the columns to the dataset
# Show the Plot
#

result = openFileDialog()

if result:
    a, b, col = result
    avg = (d[a] + d[b])/2
    diff = d[a] - d[b]

    avg.setName('average')
    diff.setName('difference')

    d.addColumn(avg)
    d.addColumn(diff)

    x = d.indexOf(avg)
    y = d.indexOf(diff)
    color = d.indexOf(col)

    showPlot(x, y, color)

#####

##### Example 2 #####
# This script will create a new sub-entitylist with those
# entities from # "All Entities" who have present "P"
# flag in at least 1 out of all samples.
#

# get the active experiment
activeExperiment = script.marray.project.getActiveExperiment()

# get the analysis object from active experiment
analysis = activeExperiment.getAnalysis()

# get the flag dataset. columns of this dataset are the flag columns from
all the samples
flagDataset =

```

```

script.marray.project.analysis.getFlagDatasetNode(analysis).getDataset()

# create int array of row indices where at least 1 of the flag columns have
"P" flag
from com.strandgenomics.cube.framework.data import DefaultIntArray
indices = DefaultIntArray()
for row in range(flagDataset.getRowCount()):
    for col in range(1, flagDataset.getColumnCount()):
        if flagDataset[col].get(row) == "P":
            indices.add(row)
            break

# get All Entities list from active experiment
allEntities = activeExperiment.getAllEntitiesEntityList()

# create new sub-entitylist
# createNewSubEntityList(parent, name, notes, indices, columns)
# where, parent = parent entitylist
#     name = name of the new entity list
#     notes = notes for the new entity list
#     indices = subset of row indices of parent entity list
#     columns = columns to add to the new entity list; size of column
should be equal to size of new entity list
#
columns = []
subEntityList =
script.marray.project.elist.createNewSubEntityList(allEntities,
"presentProbes", "notes", indices, columns)

# add new sub-entitylist to All Entities list
allEntities.addChild(subEntityList)

#*****

#***** Example 3 *****
# illustrates accessing entitylists, interpretation, samples etc.

# get active entity list of the project
entitylist = script.marray.project.getActiveProject().
    getActiveEntityList()

# get identifier column of the entity list
identifierColumn = entitylist.getIdentifierColumn()
print identifierColumn

# get all columns in the entitylist as an array
columns = entitylist.getColumns()
print columns

```

```

# get technology of entitylist
technology = entitylist.getTechnology()
print technology

# get primary mark of entity list
mark = entitylist.getPrimaryMark()
print mark

# get size of the entity list
size = entitylist.getSize()
print size

# get the active experiment
activeExperiment = script.marray.project.getActiveExperiment()

# get the array of all the samples in the active experiment
samples = activeExperiment.getSamples()

# get the number of samples
numSamples = len(samples)
print "number of samples in the active experiment: ", numSamples

*****

***** Example 4 *****
#Obtains the dataset associated with a particular entity list
#
#imports
#

from script.coercion import to_py
from com.strandgenomics.marray.project.translator import *
from com.strandgenomics.cube.framework.data import ArrayUtil
from com.strandgenomics.cube.dataset import DatasetUtil

def getIdentifierColumn(dataset): #{{{
    return to_py(dataset)[0]
# }}}

def getEntityListDataset(entityList, expt): #{{{
    analysis = exp.getAnalysis()
    ndn =
script.marray.project.analysis.getNormalisedDatasetNode(analysis)
    dataset = ndn.getDataset()
    idColumn = dataset.getColumn(0)
    eidColumn = entityList.getIdentifierColumn()
    indices = EntityListTranslator.getMatchingRowIndices(eidColumn,

```

```

        idColumn)
    indices = ArrayUtil.createIndexedIntArray(indices)
    return DatasetUtil.getRowSubsetDataset(dataset, indices)
# }}}

exp = script.marray.project.getActiveExperiment()
els = exp.getAllEntityLists()
found = 0
for el in els:
    if el.getName() == "Entitylist of selection":
        d = getEntityListDataset(el, exp)
        print "Number of rows      :", d.getRowCount()
        print "Number of columns :", d.getColumnCount()
        found = 1
        break
if found == 0:
    print "No such entity list found!"

#*****

#***** Example 5 *****
#
#Creates a dataset containing identifier, signal columns
#and annotation columns associated with a particular
#entity list and interpretation

v=script.marray.view.Table()
d=v.dataset
script.view.Table(dataset=d).show()

#*****

#***** Example 6 *****
#
# Example to illustrate the search API.
#
# imports
#

from java.util import *
from com.strandgenomics.marray.project.search import *
from com.strandgenomics.enterprise.client.filesystem import FileObject

def getObjectTypes(): #{{{
    objTypes = ArrayList()

```



```

    # Search for all objects of type Entitylist
    objTypes.add("Entitylist")

    # You can search for multiple object types simultaneously
    # objTypes.add("Sample")
    # objTypes.add("Experiment")

    return objTypes
#}}}

def getConditions(): #{{{
    conditions = ArrayList()

    # Each condition is represented using a java HashMap.
    # The HashMap has 3 keys:
    # "Search Field" : The value for this key must be a string;
    # "Search Value" : The value for this must be either a string or a java
    # Date object, as appropriate;
    # "Condition" :
    #     - If "Search Value" contains a string, then "Condition"
    # can be one of "equals", "includes", "starts with" or "ends with"
    #     - If "Search Value" contains a Date or a number,
    # then "Condition"
    # can be one of "=", "<=" or ">="
    # Please go to the Advanced Search option in Search in the Tool Bar.
    # The drop down boxes here give a list of values that are
    # applicable to each field.

    #first condition
    condition_1 = HashMap()
    condition_1.put("Search Field", "Notes")
    condition_1.put("Condition", "includes")
    condition_1.put("Search Value", "upregulated genes")

    #second condition
    condition_2 = HashMap()
    condition_2.put("Search Field", "Number of entities")
    condition_2.put("Condition", ">=")
    condition_2.put("Search Value", "30")

    #you can add more conditions if you wish.

    conditions.add(condition_1)
    conditions.add(condition_2)

    return conditions
#}}}

```

```

def getJoinByType(): #{{{
    # The conditions mentioned in getConditions can be joined using
    "OR" or # "AND"

    return "OR"
    #return "AND"
#}}}}

def getIsSimple(): #{{{
    # 0 indicates advanced search, 1 indicates simple search (no
    conditions)

    return 0
    #return 1
#}}}}

def runSearch(objTypes, conditions, joinBy, isSimple): #{{{
    searchManager = SearchHandler.getSearchManager()

    return searchManager.search(objTypes, conditions, joinBy,
isSimple)
#}}}}

def parseResults(results): #{{{
    from script.marray.eproject.search import createDataNodes

    dataset = SearchUtil.getResultsDataset(results)

    if dataset is not None:
        numRows = dataset.getRowCount()
        objectToTypeMap = LinkedHashMap()
        for index in range(0, numRows):
            objPath = dataset.getColumn("Path").get(index)
            object = FileObject(objPath)
            type = dataset.getColumn("Type").get(index)
            objectToTypeMap.put(object, type)

        dataNodes = createDataNodes(objectToTypeMap)
        return dataNodes
    else:
        print "No results found!"
        return
#}}}}

#{{{ main

objTypes = getObjectTypes()
conditions = getConditions()

```

```

joinBy = getJoinByType()
isSimple = getIsSimple()

results = runSearch(objTypes, conditions, joinBy, isSimple)
# Alternatively, for a simple search without filtering based on conditions,
you
# can try:
#
# results = runSearch(objTypes)

resultNodes = parseResults(results)


# illustrates result
for node in resultNodes:
    print node.getName()

#}}}

#*****

```

31.3 The R Editor

R scripts can be called from **GeneSpring GX** and given access to the dataset in **GeneSpring GX** via *Tools* → *Editor*. You will need to first set the path to the R executable in the Miscellaneous section of *Tools* → *Options*, then write or open an R script in this R script editor, and then click on the Run  icon. If this path is incorrect, a dialog saying “Unable to find R on this system.” will pop up.

Example R scripts are available in the `samples/rScripts` sub-folder of the installation directory; these show how the **GeneSpring GX** dataset can be accessed and sent to R for processing and how the results can be fetched back.

Help on writing R scripts in **GeneSpring GX** appears below. Broadly speaking, the input is the currently selected entity list and interpretation in the currently active experiment and the output is a sub-entity list that is automatically added as a child of the input entity list; additionally, this sub-entity list can have associated data that is generated by the R script.

31.3.1 Commands related to R input from GeneSpring GX

1. Create a data frame in R and populate it with data from **GeneSpring GX**

```

eg<-getDataset()

# Rows in data frame --> all rows corresponding to the active entity list
# Columns in data frame --> the Identifier column + all the data columns
# corresponding to the active interpretation

```

2. A data frame in R, populated with the experimental grouping information for the active interpretation

```

dg<-getExperimentGroupingDataset()

# the first column contains sample names for the active interpretation
# the second column contains condition names

```

3. Get column index

```

ci<-getColumnIndices(condition_name, eg=getExperimentGroupingDataset())

# This is with respect to the data.frame returned by getDataset( ) of
# 'condition_name' (a string value you can get by using
# expt.getConditions as shown in table 1.1

```

4. To obtain user input from a dialog

```

#user inputs, possible types are int, float, string

fdr<-askUser("FDR CutOff",float,.05)
numPerm<-askUser("Number of Permutations",int,100)

```

31.3.2 Commands related to R output to GeneSpring GX

To send results back to **GeneSpring GX**:

```

addResultColumns(data.frame(row_index, sum), c("Sum"))

#use the above function with
# 2 arguments, a dataframe and a list of output column names.
#
# Data frame first argument --> row indices from the input entity list,
# for instance if the input entity list has a 100 entities and you want
# the output to have the 1st, 10th and 15th items from the input entity
# list then make rowindex contain just 1,10,15.
#
# Data frame subsequent arguments --> any data columns associated with
# the output entity list; note that these must have exactly the same
# length as the output entity list, i.e., the length of rowindex.
#
# Output Column Names --> one name for each data column to be added to
# the output gene list

```

31.3.3 Debugging a Script

While debugging an R script, it is useful to print out the values of the important variables. This is done in R, by simply mentioning the variable name (as you would when working in R). For example, you can print out the value in sum as follows:

```

# assigning value to sum
sum <- sum[sum > 0]

# printing value in sum
sum

```

The value in the variable sum will be printed in the log window panel ('Console') of the R Editor window. Any errors occurring in R will also appear here.

Note: If an error "Unable to parse R generated output file <filename>" occurs in the Console window, it means that the execution process in R did not complete successfully and therefore **GeneSpring GX** could not retrieve a valid result.

Check that the function addResultColumns(..) has been called correctly. (The function writes data from the R program to a temporary file on the disk, which GeneSpringGX then parses to generate the child entity list)

31.3.4 Example R scripts

```
***** Example 1 *****

#The actual R script content appears below between the ‘hash lines’
#The following simple example shows how to compute the sum and
#standard deviation of the data columns
#####

# data.frame
df<-getDataset()

numrows <- nrow(df)
numcols <- ncol(df)

# first column is identifier column.
# data starts from second column.
data <- df[2:numcols]

sum <- (1:numrows)
row_index <- (1:numrows)

for (i in 1:numrows) {
  s <- sum(data[i,])
  if (s <= 0) {
    sum[i] <- -1
    row_index[i] <- -1
  }
  else {
    sum[i] <- s
    row_index[i] <- i
  }
}

sum <- sum[sum > 0]
row_index <- row_index[row_index > 0]
addResultColumns(data.frame(row_index, sum), c("Sum"))

*****

***** Example 2 *****

#-----
# R Script to run SAM in R
# Choose an entity list and an appropriate interpretation in the navigator
(currently
```

```

# working only for non-averaged interpretations)
# Run the script
# Choose an experiment parameter with 2 conditions
# Provide FDR CutOff value and Number of Permutations
# The result will be a new child entity list with SAM q-values and
regulations stored as list associated values
#-----

#user inputs, possible types are int, float, string
#-----
fdr<-askUser("FDR CutOff",float,.05)
numPerm<-askUser("Number of Permutations",int,100)

# get normalized data corresponding to the interpretation
#-----
df<-getDataset()
numrows <- nrow(df)
numcols <- ncol(df)
data_names <- names(df)

# get conditions for the chosen interpretation
#-----
eg<-getExperimentGroupingDataset()
condition_names <- unique(eg[2])
num_conditions <- length(condition_names[,1])

# The core R Script
#-----
library(Biobase)
library(samr)

# do only if number of conditions for chosen interpretation is exactly 2
if (num_conditions==2) {

  # assign class labels
  c1 <- getColumnIndices(condition_names[1,1], eg)
  c2 <- getColumnIndices(condition_names[2,1], eg)
  sample.class<-c(rep(1,length(eg[,2])))
  for (i in 1:length(c1)) {sample.class[c1[i]-1]=1}
  for (i in 1:length(c2)) {sample.class[c2[i]-1]=2}

  # set up sam
  sam.data <-
list(x=df[,2:numcols],y=sample.class,geneid=rownames(df),genenames=df[,1],lo
gged2=TRUE)
  samr.obj <- samr(sam.data, resp.type="Two class unpaired",
nperms=numPerm)

```

```

# determine delta based on fdr
delta.table <- samr.compute.delta.table(samr.obj)
delta = delta.table[1,1]
for (i in 1:nrow(delta.table)) {
  if (is.nan(delta.table[i,5]) || (fdr>delta.table[i,5]))
{delta=delta.table[i,1];break}
}

# filter on delta
siggenes.table <- samr.compute.siggenes.table(samr.obj,delta, sam.data,
delta.table)
up.genes <- siggenes.table$genes.up
down.genes <- siggenes.table$genes.lo

# set up q value and fold change vectors
rowindex <- (1:numrows)
for (i in 1:numrows) {rowindex[i] <- i}
qval <- rep('NA',numrows)
qval[as.numeric(up.genes[,3])] <- as.numeric(up.genes[,8])
qval[as.numeric(down.genes[,3])] <- as.numeric(down.genes[,8])
dir <- rep('NA',numrows)
dir[as.numeric(up.genes[,3])] <- 'UP'
dir[as.numeric(down.genes[,3])] <- 'DOWN'

# identify number of non-missing q vals, ugly, needs to be made more elegant
nonMissing=0
for (i in 1:numrows) {if (qval[i]!='NA') {nonMissing<-nonMissing+1}}
if (nonMissing>0) {
  qvalResult<-rep('NA',nonMissing)
  dirResult<-rep('NA',nonMissing)
  rowindexResult=rep(1,nonMissing)
  j=1
  for (i in 1:numrows)
    {if (qval[i]!='NA') {
      rowindexResult[j]=rowindex[i]
      qvalResult[j]=qval[i]
      dirResult[j]=dir[i]
      j=j+1
    }
  }
  # put back the entity list
  addResultColumns(data.frame(rowindexResult, qvalResult,dirResult),
c("SAMR Q-value","SAMR REGULATION"))
}
}

if (num_conditions!=2) {"The Interpretation does not have exactly 2

```


conditions"}

Chapter 32

Table of Key Bindings and Mouse Clicks

All menus and dialogs in **GeneSpring GX** adhere to standard conventions on key bindings and mouse clicks. In particular, menus can be invoked using *Alt* keys, dialogs can be disposed using the *Escape* key, etc. On Mac **GeneSpring GX** confirms to the standard native mouse clicks.

32.1 Mouse Clicks and their actions

32.1.1 Global Mouse Clicks and their actions

Mouse clicks in different views in **GeneSpring GX** perform multiple functions as detailed in the table below:

Mouse Clicks	Action
Left-Click	Brings the view in focus
Left-Click	Selects a row or column or element
Left-Click + Drag	Draws a rectangle and performs selection or zooms into the area as appropriate
Shift + Left-Click	Selects contiguous areas with last selection, where contiguity is well defined
Control + Left Click	Toggles selection in the region
Right-Click	Bring up the context specific menu

Table 32.1: Mouse Clicks and their Action

Mouse Clicks	Action
Shift + Left-Click	Draw Irregular area to select

Table 32.2: Scatter Plot Mouse Clicks

Mouse Clicks	Action
Shift + Left-Click + Move	Rotate the axes of 3D
Shift + Middle-Click + Move up and down	Zoom in and out of 3D
Shift + Right-Click + Move	Translate the axes of 3D

Table 32.3: 3D Mouse Clicks

32.1.2 Some View Specific Mouse Clicks and their Actions

32.1.3 Mouse Click Mappings for Mac

Mac Mouse Clicks	Equivalent Action in Windows/Linux
Click	Left-Click
Apple + Click	Control + Left-Click
Shift + Click	Shift + Left-Click
Control + Click	Right-Click
Alt + Click	Middle-Click

Table 32.4: Mouse Click Mappings for Mac

32.2 Key Bindings

These key bindings are effective at all times when the **GeneSpring GX** main window is in focus.

32.2.1 Global Key Bindings

Key Binding	Action
Ctrl-N	New Project
Ctrl-O	Open Project
Ctrl-X	Quit GeneSpring GX
Ctrl-F	Search on entities
Ctrl-I	Shows Entity Inspector
Ctrl-R	Brings up 'Properties' window

Table 32.5: Global Key Bindings

Bibliography

- [1] Comparison of Probe Level Algorithms. <http://affycomp.biostat.jhsph.edu>
- [2] Affymetrix Latin Square Data. http://www.affymetrix.com/support/technical/sample_data/datasets.affx
- [3] Identifying and Validating Alternative Splicing Events. <http://www.affymetrix.com/support/technical/technotes/id.altsplicingevents.technote.pdf>
- [4] Reiner A, Yekutieli D and Benjamini Y: Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics*, 19, 3, (368-375), 2003.
- [5] Alternative Transcript Analysis Methods for Exon Arrays. <http://www.affymetrix.com/support/technical/whitepapers/exon.alt.transcript.analysis.whitepaper.pdf>
- [6] Belsley DA, Kuh E, and Welsch RE, *Regression Diagnostics*. Hoboken, NJ: John Wiley & Sons, Inc., 1980.
- [7] Benjamini B and Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* 57, 289-300, 1995.
- [8] The Bioconductor Webpage. <http://www.bioconductor.org>.
Validation of Sequence-Optimized 70 Base Oligonucleotides for Use on DNA Microarrays, Poster at <http://www.operon.com/arrays/poster.php>.
- [9] Joshua M Korn, *et al.*: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare NVs, *Nature Genetics*, 40, 10, 2008 <http://www.nature.com/ng/journal/v40/n10/abs/ng.237.html>
- [10] Bolstad, BM, Irizarry RA, Astrand M, and Speed, TP: A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193 Supplemental information, 2003.
- [11] Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 2, 185-193, 2003.
- [12] BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set <http://www.affymetrix.com/support/technical/whitepapers/brlmm.whitepaper.pdf>

- [13] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, and Nickerson DA: Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium, *Am. J. Hum. Genet.*, 74, 106120, 2004. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1181897>
- [14] Olshen AB, Venkatraman ES, Lucito R, Wigler M.: Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, Oct;5(4):557-72, 2004. <http://www.ncbi.nlm.nih.gov/pubmed/15475419>
- [15] Adam B. Olshen, Change-Point Analysis of Microarray-Based DNA Copy Number Data. <http://www.mskcc.org/mskcc/shared/graphics/epidemiology/AdamOlshen/Template.pdf>
- [16] DNA Copy Number Data Analysis. <http://bioconductor.org/packages/2.4/bioc/html/DNAcopy.html>
- [17] DChip: The DNA Chip Analyzer. <http://www.biostat.harvard.edu/complab/dchip>.
- [18] Dempster AP, Laird, NM, and Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* 39, 1-38 (1977).
- [19] Devlin B and Risch N: A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping, *Genomics*, 29, 311322, 1995.
- [20] Dobson AJ. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.
- [21] Dudoit S, Yang H, Callow MJ, and Speed TP: Statistical Methods for identifying genes with differential expression in replicated cDNA experiments, *Stat. Sin.* 12, 1, 11-139, 2000.
- [22] Gabriel S B, *et al.*: The Structure of Haplotype Blocks in the Human Genome, *Science*, Vol. 296., No. 5576, 2225 - 2229, 2002. <http://www.sciencemag.org/cgi/content/abstract/296/5576/2225>
- [23] GISTIC documentation <http://www.broad.mit.edu/cancer/pub/GISTIC/>
- [24] Glantz S: *Primer of Biostatistics*, 5th edition, McGraw-Hill, 2002.
- [25] GeneLogic Latin Square Data. <http://qolotus02.genelogic.com>.
- [26] GeneLogic Spike In Study. <http://www.genelogic.com/media/studies/spikein.cfm>
- [27] The International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs, *Nature*, October 18; 449(7164): 851-861, 2007 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689609/?tool=pubmed>
- [28] Hill AA, Brown EL, Whitley MZ, Tucker-Kellog G, Hunter CP, and Slonim DK: Evaluation of normalization procedures for Oligonucleotide array data based on spiked cRNA controls, *Genome Biology*, 2, 0055.1-0055.13, 2001.
- [29] Hoffmann R, Seidl T, and Dugas M: Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis, *Genome Biology*. 3(7), 0033.1-0033.11, 2002.
- [30] Hubbell E, *et al.*: Robust estimators for expression analysis. *Bioinformatics*. 18(12):1585-92, 2002.

- [31] Hubbell, E., Designing Estimators for Low Level Expression Analysis. <http://mbi.osu.edu/2004/ws1abstracts.html>
- [32] Irizarry, RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed T.P: Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 4(2), 249-264, 2003.
- [33] Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed: Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15, 2003.
- [34] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP: Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* .Vol. 4, Number 2: 249-264, [Abstract, PDF, PS, Complementary Color Figures-PDF, Software], 2003.
- [35] Lewontin RC. On measures of gametic disequilibrium. *Genetics* 120: 849852, 1988
- [36] Li C and Wong WH: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA*. 98, 31-36, 2000.
- [37] Li C and Wong WH: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology*. 2(8), 0032.1-0032.11, 2001.
- [38] Li C and Wong WH: Model based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *PNAS* Vol. 98: 31-36, 2001.
- [39] The Lowess method. <http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm>.
- [40] Statistical Algorithms Description Document, Affymetrix Inc. <http://www.affymetrix.com/support/technical/whitepapers/sadd.whitepaper.pdf>.
- [41] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.
- [42] Pritchard JK and Przeworski M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69: 1-14. 2001.
- [43] Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* 77, 2, Feb 1989. <http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf>.
- [44] Reich DE and Goldstein DB: Detecting Association in a Case-Control Study While Correcting for Population Stratification, *Genetic Epidemiology* 20:416, 2001.
- [45] Shaw RG and Olds TM: ANOVA for Unbalanced Data: An overview, *Ecology*, 74, 6, (1638-1645), 1993.
- [46] Speed T: Always log spot intensities and ratios, Speed Group Microarray Page. <http://stat-www.berkeley.edu/users/terry/zarray/Html/log.html>.
- [47] Speed FM, Hocking RR, and Hackney OP: Methods of Analysis of Linear Models with Unbalanced Data, *J. Am Stat Assoc*, 73, 361, (105-112), 1978.

- [48] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D: Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, 38 (8), 904-909, 2006. <http://www.nature.com/ng/journal/v38/n8/abs/ng1847.html>
- [49] Patterson N, Price AL, and Reich D: Population Structure and Eigenanalysis, *PLoS Genet* 2(12): e190. doi:10.1371/journal.pgen.0020190 <http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020190>
- [50] Strand Life Sciences **GeneSpring GX**. <http://avadis.strandls.com>
- [51] Westfall PH, Young SS: Resampling based multiple testing. John Wiley and Sons. New York, 1993.
- [52] Wu Z, Irizarry RA, Gentleman R, Murillo FM, and Spencer F: A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 1 (May 28, 2004).
- [53] Benjamini Y, and Yekutieli D: The control of false discovery rate under dependency, *Ann Stat*, 29, (1165-1188), 2001.
- [54] Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, and Ehm MG Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals *Hum Hered* 2002;53:7991

Index

- χ^2 correlation test, [787](#)
- χ^2 test, [785](#)
- activation, [4](#), [7](#), [10](#)
- advanced workflow
 - create gene level experiment, [502](#)
 - create interpretation, [499](#)
 - experiment grouping, [498](#)
 - experiment setup, [497](#)
- affymetrix, [185](#)
- agilent mirna, [381](#)
 - sample validation, [387](#)
- annotation
 - Gene Ontology Browsing, [609](#)
- ANOVA
 - F-ratio, [552](#)
 - Friedman, [554](#)
 - Kruskal-Wallis, [553](#)
 - N-way, [555](#)
 - N-way
 - type-III ss, [555](#)
 - one-way, [551](#)
 - Repeated Measures, [554](#)
 - sum of square, [551](#)
 - Welch, [553](#)
- association
 - views, [798](#)
- association analysis, [761](#)
 - χ^2 correlation test, [787](#)
 - χ^2 test, [785](#)
 - analysis, [777](#)
 - Birdseed Report, [767](#)
 - cochran-armitage test, [787](#)
 - EIGENSTRAT correction, [777](#)
 - EIGENSTRAT filter, [768](#)
 - filter samples by missing values, [766](#)
 - filter snps by hwe p-value, [775](#)
 - filter snps by maf, [776](#)
 - filter snps by minor allele frequency, [776](#)
 - filter snps by missing value, [772](#)
 - filters, [772](#)
 - find significant snps, [792](#)
 - fisher's exact test, [786](#)
 - genome browser, [798](#)
 - genomic control, [784](#)
 - haplo block, [132](#), [795](#)
 - haplotypes, [793](#)
 - haplotypes view, [132](#), [795](#)
 - haplotyping trend regression, [793](#)
 - hardy-weinberg equilibrium, [775](#)
 - Identify overlapping genes, [800](#)
 - identify snps with differential missingness, [773](#)
 - ld analysis, [796](#)
 - linkage disequilibrium, [796](#)
 - mlr, [793](#)
 - mode of inheritance, [782](#)
 - mtc, [784](#)
 - multiple linear regression, [793](#)
 - multiple logistic regression, [793](#)
 - multiple testing correction, [784](#)
 - permutative p-value computation, [784](#)
 - qc, [765](#)
 - quality control, [765](#)
 - results analysis, [800](#)
 - snp regression, [792](#)
 - snp tagging, [790](#)
 - statistical analysis, [782](#)
 - technology, [762](#)
 - utilities, [800](#)
- classification, [585](#)
 - build prediction model, [588](#)
 - confusion matrix, [605](#)
 - decision trees, [591](#)
 - lorenz curve, [606](#)
 - naive bayesian, [602](#)

- neural network, [596](#)
- pipeline, [586](#)
- plsd, [603](#)
- predicting outcomes, [605](#)
- report, [606](#)
- support vector machines, [599](#)
- training, [588](#)
- validation, [586](#)
- viewing results, [605](#)
- classification report, [606](#)
- classify
 - decision trees, [605](#)
 - neural network, [605](#)
 - SVM, [605](#)
- cluster set, [567](#)
- clustering, [563](#)
 - classification, [564](#)
 - cluster set, [567](#)
 - combined tree, [564](#)
 - condition trees, [564](#)
 - dendrogram, [570](#)
 - distance measures, [578](#)
 - gene tree, [564](#)
 - graphical views, [566](#)
 - hierarchical, [580](#)
 - k-means, [580](#)
 - missing values, [583](#)
 - pipelines, [564](#)
 - self organizing maps, [582](#)
 - U matrix, [577](#)
 - what is, [563](#)
- cochran-armitage test, [787](#)
- confusion matrix, [605](#)
- copynumber, [707](#)
 - affybatcheffectcorrection, [721](#)
 - affyworkflow, [717](#)
 - algorithms, [749](#)
 - analysis, [727](#)
 - configuration, [747](#)
 - create custom reference, [745](#)
 - disc cache, [747](#)
 - filters, [733](#)
 - gistic, [729](#)
 - gistic overlapping genes, [733](#)
 - illumina, [743](#)
 - performance statistics, [749](#)
 - resultsanalysis, [741](#)
 - technology, [708](#)
 - terminology, [708](#)
 - translation, [747](#)
 - tutorials, [760](#)
 - utilities, [741](#)
 - views, [739](#)
- custom affy using cdf, [181](#)
- custom agilent arrays, [378](#)
- custom illumina technology, [299](#)
- decision trees, [591](#)
 - classify, [605](#)
 - model, [595](#)
 - training, [594](#)
- dendrogram, [570](#)
- Differential Expression Analysis, [545](#)
- distance measures, [578](#)
 - chebychev, [579](#)
 - differential, [579](#)
 - euclidean, [578](#)
 - manhattan, [579](#)
 - pearson absolute, [579](#)
 - pearson centered, [579](#)
 - pearsons uncentered, [579](#)
 - squared euclidean, [578](#)
- EIGENSTRAT correction, [777](#)
- entitylist, [27](#)
- es value, [625](#), [633](#)
- experiment, [22](#)
- find significant haplotypes, [793](#)
- find significant snps, [792](#)
- fisher's exact test, [786](#)
- generic single color, [433](#)
- generic two color, [461](#)
- genes et, [623](#), [631](#)
- genome browser, [798](#), [803](#)
- genomebrowser
 - copy number experiments, [819](#)
 - faq, [821](#)
 - import tracks, [809](#)
 - track properties, [817](#)
 - trackoperations, [817](#)
 - tracks, [804](#)
 - working, [807](#)
- genomebrowser.visualization, [805](#)
- genomic control, [784](#)
- geoimport, [489](#)

- experiment parameters, [491](#)
 - load dataset, [489](#)
 - possible error messages, [492](#)
- getting started
 - config, [40](#)
 - help, [41](#)
 - script, [40](#)
 - ui, [17](#)
- GO Analysis, [611](#)
- GO Terms, [609](#)
- goanalysis, [609](#)
- graphical views
 - clustering, [566](#)
- groups, [849](#)
- gsa, [631](#)
 - genes et, [631](#)
 - introduction, [631](#)
- gsea, [623](#)
 - genes et, [623](#)
 - introduction, [623](#)
- gswgclient, [849](#)
- guided workflow, [141](#), [193](#), [265](#), [303](#), [343](#), [383](#)
 - affymetrix, [145](#)
 - agilent mirna, [388](#)
 - agilent single color, [309](#)
 - agilent two color, [348](#)
 - exonexpression, [197](#)
 - illumina, [268](#)
- haplo block, [132](#), [795](#)
- haplotypes, [793](#)
- haplotypes view, [132](#), [795](#)
- haplotyping trend regression, [793](#)
- hardy-weinberg equilibrium, [775](#)
- hierarchical, [580](#)
- identify snps with differential missingness, [773](#)
- inheriting permissions, [854](#)
- installation, [1](#)
 - activation, [4](#), [7](#), [10](#)
 - copy number and association experiments, [2](#)
 - Linux, [5](#)
 - mac, [8](#)
 - upgrade, [15](#)
 - windows, [2](#)
- interpreting task logs, [857](#)
- ipa, [823](#)
- k-means, [580](#)
- ld analysis, [796](#)
- license manager, [12](#)
- linkage disequilibrium, [796](#)
- Linux
 - install, [6](#)
 - uninstall, [8](#)
- login dialog, [850](#)
- lorenz curve, [606](#)
- Mac
 - uninstall, [12](#)
- mac
 - install, [9](#)
- Mann-Whitney, [550](#)
- Mann-Whitney
 - paired, [550](#)
- migration, [61](#)
- minor allele frequency, [776](#)
- mlr, [793](#)
- mode of inheritance, [782](#)
- model
 - decision trees, [595](#)
 - naive bayesian, [603](#)
 - neural network, [597](#)
 - SVM, [601](#)
- multiple linear regression, [793](#)
- multiple logistic regression, [793](#)
- Multiple Testing Correction, [557](#)
- Multiple Testing Correction
 - Benjamini-Hochberg method, [559](#)
 - Benjamini-Yekutieli method, [560](#)
 - Bonferroni, [558](#)
 - Holm method, [558](#)
 - Westfall-Young method, [559](#)
- naive bayesian, [602](#)
 - model, [603](#)
 - train, [602](#)
- nes value, [625](#), [633](#)
- neural network, [596](#)
 - classify, [605](#)
 - model, [597](#)
 - training, [597](#)
- Normalization
 - Lowess, [548](#)
 - Quantile, [546](#)
- object ownership, [851](#)
- object permissions, [851](#)

- objects, [850](#)
- pathway, [639](#)
- permission conflicts, [852](#)
- permutative p-value computation, [784](#)
- pipelines
 - classification, [586](#)
 - clustering, [564](#)
- platforms, [1](#)
- Post-Hoc, [552](#)
- Post-Hoc
 - SNK, [553](#)
 - Tukey HSD, [552](#)
- prediction
 - what is, [585](#)
- project, [21](#)
- propagating permissions, [852](#)
- remote execution, [854](#)
 - add or remove samples, [856](#)
 - clustering, [856](#)
 - experiment creation, [855](#)
 - snp tagging, [856](#)
- rtpcr, [421](#)
- sample, [23](#)
- scatterplot
 - visualization, [88](#)
- scripts, [859](#)
- self organizing maps, [582](#)
- servermigration, [66](#)
- significant haplotypes, [793](#)
- significant snps, [792](#)
- snp regression, [792](#)
- SOM, [582](#)
 - U matrix, [577](#)
- summarization, [185](#)
- summarization
 - GCRMA, [187](#)
 - Invariant Set, [187](#)
 - Li-Wong, [187](#)
 - MAS5
 - Absolute Calls, [189](#)
 - median polish, [187](#)
 - PLIER, [188](#)
 - quantile, [186](#)
 - RMA, [186](#)
 - Tukey-BiWeight, [188](#)
- support vector machines, [599](#)
 - SVM, [599](#)
 - classify, [605](#)
 - model, [601](#)
 - training, [600](#)
- t-test, [549](#)
- t-test
 - paired, [549](#)
 - Welch, [550](#)
- tag snps, [790](#)
- task manager, [854](#)
- technology
 - association analysis, [762](#)
- Thresholding, [545](#)
- train
 - naive bayesian, [602](#)
- training
 - classification, [588](#)
 - decision trees, [594](#)
 - neural network, [597](#)
 - SVM, [600](#)
- U matrix, [577](#)
- ui, [17](#)
 - desktop, [18](#)
 - desktop navigator, [19](#)
 - legend, [20](#)
 - workflow, [19](#)
- update, [41](#)
- users, [849](#)
- validation
 - for classification, [586](#)
- visualization, [73](#)
 - haplo block, [132](#)
 - haplotypes view, [132](#)
- windows
 - install, [3](#)
 - uninstall, [5](#)