Next generation sequencing guide

> SEE WHAT MORE WE CAN DO FOR **YOU** AT **WWW.IDTDNA.COM**.

INTEGRATED DNA TECHNOLOG

custom oligos • next generation sequencing • CRISPR genome editing • qPCR & PCR • synthetic biology • functional genomics

For Research Use Only | Version 1

Table of contents

1. Intro to next generation sequencing (NGS)

NGS, or high-throughput sequencing, enables sequence profiling of everything from genomes and transcriptomes to DNA-protein interactions. These technologies are an integral part of genetic research and discovery. The ability to generate large amounts of sequence data in a relatively short amount of time enables a wide range of genetic analysis applications and accelerates advances in research, clinical, and applied markets.

2. Recent history of NGS

What started with the first generation of sequencing, Sanger sequencing, exploded into a genomics revolution. After the completion of the Human Genome Project, the next generation of sequencing gave rise to several major methods of sequencing.

The 454 Sequencing® method (Roche) was introduced in 2005 and uses pyrosequencing based on sequencing by synthesis. First, the strand to be sequenced is replicated. As the DNA incorporates each nucleic acid, a pyrophosphate is released, giving off light that is detected and used to identify the nucleic acid added. Roche 454 technology produces sequences with the longest read length.

Supported Oligo Ligation Detection (SOLiD™) sequencing offered by Thermo Fisher Scientific was introduced in 2006. This method uses sequencing by ligation followed by emulsion PCR template preparation. The oligo probes used during ligation are fluorescently tagged. When a sequence match occurs, the resulting fluorescence indicates sequence information. SOLiD sequencing has the lowest false positive rate among whole genome sequencing platforms [1].

Another type of sequencing also introduced in 2006 is now offered by Illumina. The first Solexa sequencer was called the Genome Analyzer and used the sequencing-by-synthesis method. Fluorescently tagged bases emit a signal as they are added to a nucleic acid chain during synthesis. Illumina sequencing provided the highest throughput, largest data output, and lowest reagent costs because it could sequence 1 gigabase of data in a single run.

In 2010, ion semiconductor sequencing was released, developed by Ion Torrent Systems Inc. This technology uses an ion sensor to measure the release of hydrogen ions as nucleotides are incorporated during synthesis. The low-cost equipment is compact, and sequencing can be performed quickly.

In 2011, Pacific Biosciences (PacBio) released Single Molecule, Real-Time (SMRT™) sequencing. SMRT sequencing uses zero-mode waveguides to isolate single molecules of DNA. Like Illumina sequencing, fluorescent labeling is used to read the sequence during synthesis. While these methods are no longer being used, they paved the way for next generation sequencing methods that we use today. The SMRT sequencing was the basis for PacBio's development of long-read (more than 300 bases) sequencing and measuring DNA modifications.

One of the newest sequencing methods is nanopore sequencing by Oxford Nanopore. Nanopore sequencing uses electric current to identify sequences. Use of a nanopore allows much longer reads to be sequenced. As each nucleic acid passes through the nanopore, its electric charge is detected and recorded to determine the sequence. There are several methods of nanopore sequencing, both biological and solid state, including a version which incorporates fluorescence.

The major advantages of all these methods is the ability to sequence in parallel. Sequencing multiple reads simultaneously dramatically reduced time and cost associated with sequencing and increased the coverage quality and data output. Of these platforms, the Illumina sequencing platform is the most widely used, so the rest of this guide will focus on Illumina methods.

3. Steps in sequencing

The steps in sequencing can vary greatly depending on the type of sequencing performed. The following section describes steps used for short-read sequencing on Illumina instruments, as they are more applicable to a wider variety of sequencing applications.

3.1 Sample preparation

3.1.1 Input mass

The amount of sample you have is an important factor in determining the types of sequencing available to you. Processing steps designed to increase the sample amount, like PCR amplification, can introduce biases. Other processing steps like purification and converting to cDNA can reduce your sample amount, limiting your sequencing choices.

3.1.2 Sample type

The quality of your samples can impact your sequencing results. Samples derived from cell lines typically produce more robust sequencing outcomes. Formalin-fixed, paraffin-embedded (FFPE) tissue typically is lower quality and requires extra steps to extract the sample DNA or RNA before library preparation. FFPE quality can vary dramatically, and the extracted sample is often damaged. Sample extraction includes paraffin removal, heat exposure to remove the crosslinking, and either column- or magnetic bead-based purification, followed by quality control to ensure the library preparation success. Many researchers overcome the difficulties of working with degraded samples by starting with more sample for library preparation.

3.2 Library preparation

Before DNA or RNA samples can be sequenced, they must be fragmented, end repaired, and collected into adapter-ligated libraries. Library preparation protocols can influence the results generated by your NGS experiment. The major steps of library preparation are summarized below:

3.2.1 RNA

RNA is single-stranded and less stable than double-stranded DNA. To protect the fidelity of the RNA, it is usually converted to cDNA before forming libraries. cDNA is complementary DNA, a copy of the RNA. Recently, nanopore sequencing has enabled sequencing RNA directly, without conversion to cDNA, but this method can only be run on a few samples at a time, so it is not feasible for high-throughput applications.

3.2.2 Fragmentation and end repair

Short-read sequencing technologies like those from Illumina, cannot readily analyze very long DNA strands, so samples are fragmented into uniform pieces to make them amenable to sequencing. Samples are normally fragmented using enzymatic or mechanical methods, such as ultrasonication (Covaris). After fragmentation, the DNA fragments are end repaired or end polished. Generally, a single adenine base is added to form an overhang via an A-tailing reaction. This A overhang allows adapters containing a single thymine overhanging base to base pair with the DNA fragments.

3.2.3 Addition of adapters

A ligase enzyme covalently links the adapter and insert DNA fragments, making a complete library molecule. These adapters serve multiple functions. They contain P5 and P7 sites which enables attachment of the library molecule to the flow cell and binding sites for sequencing primers. They contain at least one barcode, also called an index, to identify samples and permit multiplexing (Figure 1). Read more about adapters for next generation sequencing.

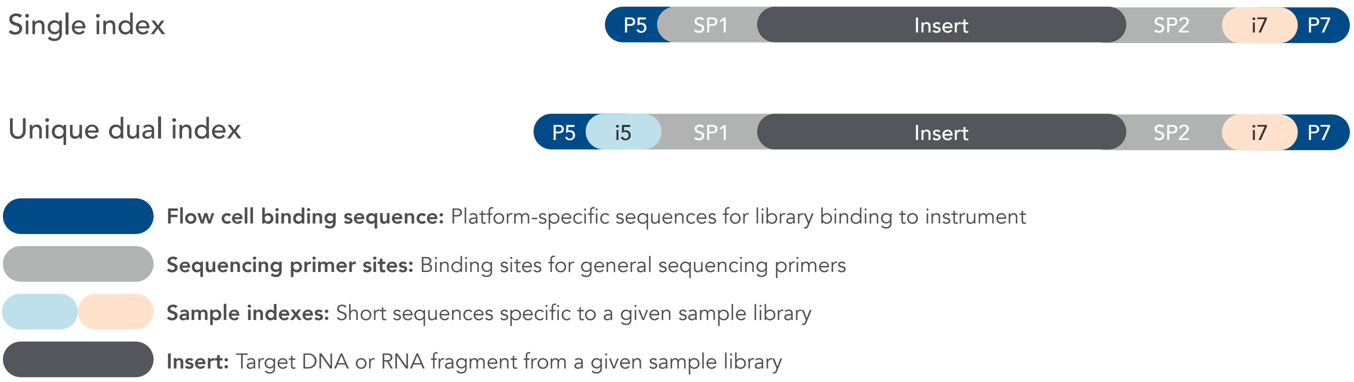
Single index
P5
SP1
Insert
SP2
i7
P7
Unique dual index
P5
i5
SP1
Insert
SP2
i7
P7
Flow cell binding sequence: Platform-specific sequences for library binding to instrument
Sequencing primer sites: Binding sites for general sequencing primers
Sample indexes: Short sequences specific to a given sample library
Insert: Target DNA or RNA fragment from a given sample library

Figure 1. Next generation sequencing adapter designs.

3.2.4 PCR amplification (optional) and clean up

Whether or not you amplify your libraries depends on the adapter type and DNA input used. At the end of library prep, whether or not you amplify with PCR, remaining oligonucleotides and small fragments must be removed. PCR clean-up can be performed using magnetic beads or a spin column.

3.3. Enrichment

Enrichment is a type of selection used for targeted sequencing. If you are preparing your samples for targeted sequencing, they must be enriched before the sequencing step. Some targeted sequencing methods enrich by selecting sequences using hybrid capture. Alternatively, amplicon library preparation methods enrich by amplifying specific sequences to create libraries using one workflow. Enrichment methods are discussed in more detail in Section 4.1.2 Targeted sequencing.

4. Types of sequencing and their applications

4.1 DNA

4.1.1 Whole genome sequencing (WGS)

Whole genome sequencing is used to initially determine the genome sequence of an organism. Additionally, WGS can be used to determine variant (mutation) frequencies within populations of organisms and to associate genetic variants with disease through genome-wide association studies (GWAS). A GWAS performs WGS on two populations and compares trait differences with genetic differences to associate identified traits with identified variants. WGS was first used for clinical diagnostics in 2009, but time and costs have limited its use in this area [1,2]. As the price of WGS decreases, it is becoming more common as a diagnostic tool. Having achieved the "$1000 genome" [4], multiple companies are pushing towards the next goal of the "$100 genome" [5,6].

4.1.2 Targeted sequencing

Targeted NGS allows users to sequence specific areas of the genome for in-depth analyses in a more rapid, cost-effective way than whole genome sequencing (Figure 1). Targeted sequencing detects known and novel variants within your region of interest. This method generally produces a smaller amount of data than WGS, making analysis more manageable.

There are several methods of targeted sequencing, each appropriate for specific applications. The most popular methods are hybridization capture, amplicon sequencing, and molecular inversion probes (MIPs). For a more in-depth comparison of hybridization capture and amplicon sequencing, see our Targeted sequencing guide.

4.1.2.1 Whole exome sequencing (WES)

Whole exome sequencing identifies all the protein-coding genes in the genome. Focusing on protein-coding exons (and excluding other regions of the genome) can lower the cost and time of sequencing, as exons make up only 1% of the genome. Variants in protein-coding exons are responsible for many diseases, so this level of sequencing is often sufficient for diagnostic applications. WES is a more practical method for mapping variants that are rare in the population to elucidate complex disorders [7]. It is also a feasible option for discovery science [8]. WES is particularly useful in oncology research and is currently used for cancer diagnostics [9]. Information gained from WES can provide insight into prognoses and personalized treatment options [10]. WES is most often carried out with hybridization probes rather than amplicons.

4.1.2.2 Hybridization capture

Prior to hybridization capture, samples are converted into sequencing libraries. Regions of interest in this library are then captured using long oligonucleotide biotinylated baits (Figure 2). Because the DNA was random sheared during library preparation, captured fragments are overlapping and unique. Baits can be tiled, overlapped, and positioned to overcome challenges of repetitive sequences, etc. With advanced design, capture can be made very uniform. Hybridization capture is often used for targeted exome sequencing. Other applications include genotyping, rare variant detection, and oncology diagnostics [11].

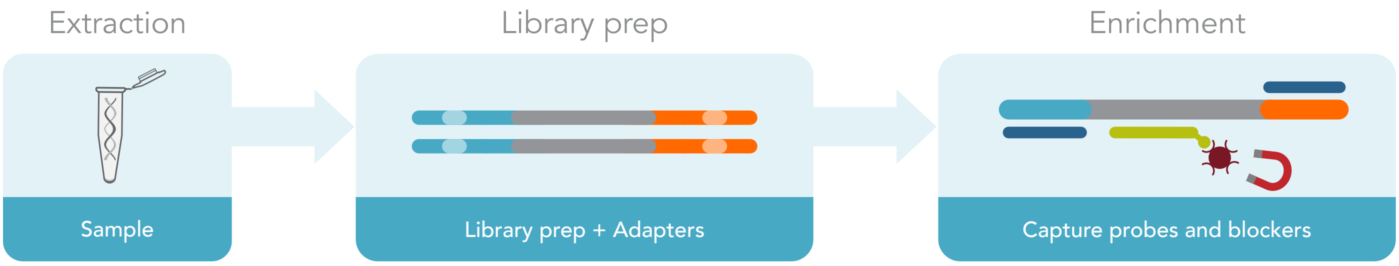

Figure 2. Hybridization capture workflow.

4.1.2.3 Amplicon sequencing

Amplicon sequencing is a highly targeted approach that enables you to analyze genetic variation in specific genomic regions. This method uses PCR to amplify DNA to make amplicons. The amplicons are indexed and sequenced (Figure 3). Amplicon sequencing is typically used for disease-associated variant detection and diagnostics [12]. It can also be used for genotyping by sequencing and to confirm CRISPR genome edits. Read more about how you can evaluate CRISPR-Cas9 edits quickly and accurately with rhAmpSeq targeted sequencing.

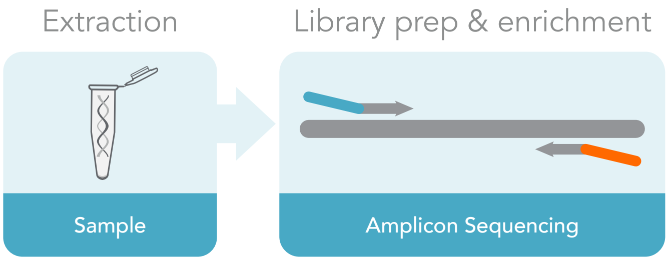

Figure 3. Amplicon sequencing workflow.

4.1.2.4 Molecular inversion probes (MIPs)

Molecular inversion probes are another common target enrichment method. Target-specific sequences are ligated to both ends of a universal sequence to make the MIP. The MIP hybridizes to the region of interest before a gap-filling reaction and a second ligation closes the circles. A restriction enzyme in the MIP may be used to create a linear molecule. The target sequences are amplified before sequencing (Figure 4). MIPs are particularly useful for large-scale genotyping.

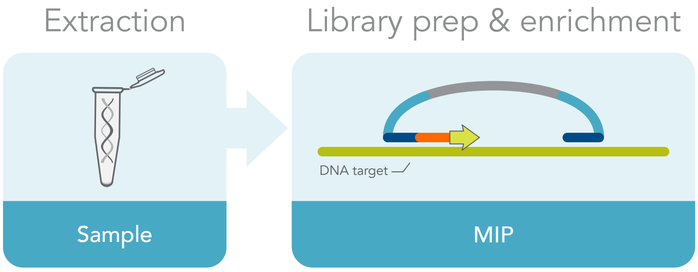

Figure 4. Molecular inversion probe workflow.

4.2. RNA

4.2.1. Whole transcriptome sequencing (WTS)

Since RNA expression level can vary based on cell type and disease state, understanding the transcriptome can provide valuable insight. Sequencing the whole transcriptome provides the most comprehensive data because it contains information regarding both coding and noncoding RNA and both known and novel variants. A variation of WTS, or RNA-seq, is stranded RNA-seq, which retains information about the strand of origin of transcripts. This makes it possible to identify novel transcripts, including antisense RNAs. RNA-seq is often used for discovery science. Typically, RNA-seq is used to evaluate messenger RNA (mRNA) expression levels and can also be used to evaluate changes in gene expression over time, gene fusions, single nucleotide polymorphisms (SNPs), alternative splicing, and RNA modifications.

4.2.2. Targeted gene expression with RNA-seq

Like DNA, RNA can be targeted for sequencing using hybridization capture or amplicon sequencing technologies. Specific populations of RNA can be targeted. Most often, coding transcripts are targeted, but other populations of RNA such as tRNA (transfer RNA), miRNA (microRNA) or small RNA may be enriched.

4.2.2.1 Ribosomal RNA depletion

Ribosomal RNA (rRNA) makes up 80–95% of the cell's RNA; however, since its expression is constant, it is rarely of interest. Therefore, it can be advantageous to avoid sequencing these molecules and focus your sequencing on more useful data. rRNA may be removed from samples by 1 of 2 methods: (1) Biotinylated probes can be used to bind the rRNA and thus remove it from the RNA sample, (2) DNA probes that bind the rRNA can be used in conjunction with RNase H to degrade the rRNA from the sample before library prep.

4.3 Epigenomics

4.3.1. ChIP-Seq

Chromatin immunoprecipitation (ChIP) is used to evaluate protein interactions with DNA. Protein can regulate DNA, impacting its expression. This type of regulation can be influenced by the environment and can change over time. ChIP identifies sites in the DNA sequence where protein is bound. An antibody is used to bind the protein of interest, which allows immunoprecipitation of the DNA bound to the protein. When this type of identification is done using an array, it is called ChIP-chip. Evaluating protein interactions of the whole genome by sequencing is called ChIP-Seq. ChIP-Seq experiments usually focus on transcription factors, histones, and histone modifications so they can reveal information about gene regulation, cell proliferation, and disease progression.

4.3.2. Methyl-Seq

Methylation, another method by which DNA is regulated, is also influenced by the environment and can change over time. Methyl groups are added to the DNA sequence and can repress DNA expression. Methyl-Seq, also known as bisulfite sequencing, treats the DNA with bisulfite before sequencing to provide information about the methylation status of the DNA sequence. This information is primarily used to evaluate gene-environment interactions.

5. Analysis

There are several steps in the data analysis process. Generally, reads first go through quality control. They may need to be demultiplexed or have adapter information removed. Next, the reads are aligned to a reference genome, transcriptome, or epigenome. If a reference genome is not available, as may be the case for less studied organisms, the sample genome may need to be assembled from the experimental reads. Finally, the sequencing data can be evaluated for variants, repeats, and other significant characteristics.

6. Follow-up experiments

Interpreting data gleaned from sequencing can be the end goal. Sequencing information is increasingly being used for clinical diagnostics and can help physicians choose personalized treatment options for their patients. Going further, evaluating how patients with specific genetic sequences respond to personalized treatment can help optimize their treatment.

In the basic science lab, sequencing information can be used to understand gene function. Genes with abnormalities can be knocked out or knocked in to cells or model organisms. Scientists are working on manipulating genes in humans for personalized genomic solutions.

7. Future directions of NGS: New platforms and applications

Methods to sequence single-cell DNA or RNA and single molecules are rapidly evolving. One of the limitations of NGS is the length of reads. Reads sequenced on the Illumina platform typically cannot exceed a few hundred base pairs. Therefore, analysis includes assembling reads and aligning them to a reference genome. Other sequencing methods like PacBio and nanopore technologies are able to sequence long reads, so the additional steps of aligning and assembling can be optional. However, these methods often require more sample input and result in lower read quality. Both short- and long-read platforms can work in conjunction to provide researchers the most flexibility and best options to increase their discovery power.

8. References

1. Rieber N, Zapatka M, et al. (2013) Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One 8(6):e66621.
2. Welch JS, Westervelt P, et al. (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. Jama 305(15):1577–1584.
3. Lunshof JE, Bobe J, et al. (2010) Personal genomes in progress: from the human genome project to the personal genome project. Dialogues Clin Neurosci 12(1):47–60.
4. Hayden EC (2014) Technology: The $1,000 genome. Nature 507(7492):294–295.
5. Herper M (2017) Illumina promises to sequence human genome for $100—but not quite yet. Forbes www. forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet. Accessed November 5, 2019.
6. McMorrow D (2010) The $100 genome: Implications for the dod, MITRE CORP MCLEAN VA JASON PROGRAM OFFICE.
7. Williams HJ, Hurst JR, et al. (2016) The use of whole-exome sequencing to disentangle complex phenotypes. Eur J Hum Genet 24(2):298–301.
8. Bamshad MJ, Ng SB, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12(11):745–755.
9. Kamps R, Brandao RD, et al. (2017) Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. Int J Mol Sci 18(2):308.
10. Rabbani B, Nakaoka H, et al. (2016) Next generation sequencing: implications in personalized medicine and pharmacogenomics. Mol Biosyst 12(6):1818–1830.
11. Necchi A, Bratslavsky G, et al. (2019) Genomic features for therapeutic insights of chemotherapy-resistant, primary mediastinal nonseminomatous germ cell tumors and comparison with gonadal counterpart. Oncologist 24(4):e142–e145.
12. Shin S, Kim Y, et al. (2017) Validation and optimization of the Ion Torrent S5 XL sequencer and Oncomine workflow for BRCA1 and BRCA2 genetic testing. Oncotarget 8(21):34858–34866.

Next generation sequencing guide

Technical support: applicationsupport@idtdna.com

> For more than 30 years, IDT's innovative tools and solutions for genomics applications have been driving advances that inspire scientists to dream big and achieve their next breakthroughs. IDT develops, manufactures, and markets nucleic acid products that support the life sciences industry in the areas of academic and commercial research, agriculture, medical diagnostics, and pharmaceutical development. We have a global reach with personalized customer service.

> SEE WHAT MORE WE CAN DO FOR **YOU** AT **WWW.IDTDNA.COM**.

For Research Use Only.

© 2020 Integrated DNA Technologies, Inc. All rights reserved. rhAmpSeq is a trademark of Integrated DNA Technologies, Inc. and is registered in the USA. All other marks are the property of their respective owners. For specific trademark and licensing information, see www.idtdna.com/trademarks. 4850 1020

INTEGRATED DNA TECHNOLOGIES