An Introduction to Next-Generation Sequencing for Cardiology

www.illumina.com/technology/next-generation-sequencing.ilmn

illumina®

Table of Contents

Part I. Welcome to Next-Generation Sequencing

NGS in Clinical Cardiology Research

In 1990, the beta cardiac myosin heavy chain gene (*MYH7*) was first implicated in hypertrophic cardiomyopathy (HCM). Since then, many genes have been associated with channelopathies, cardiomyopathies, and aortopathies, and investigators continue to look to the human genome for answers. Researchers can achieve 60% detection of causative variants for certain cardiac conditions (e.g., LQTS, HCM, ACM/ARVC, CPVT) and genetic testing may reveal whether family members are carriers or at high risk for sudden cardiac arrest.[1] Recent advances in technology are removing many of the barriers that have traditionally made genetic testing difficult.

The value of genetic testing is gaining recognition among leading organizations. For example, in 2011 the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA) published guidelines on the state of genetic testing for the channelopathies and cardiomyopathies.[1]

The introduction of next-generation sequencing (NGS) technology has transformed the way scientists and clinical researchers think about genetic information. NGS offers several advantages over existing methods, such as PCR and Sanger sequencing. Because NGS can assess multiple genes in a single assay, it eliminates the need for multiple single-gene tests to identify the causative variant or variants. With NGS, investigators can screen more than 100 clinically relevant genes simultaneously while generating highly sensitive and specific results, increasing the ability to identify causal variants. By reducing costs and shortening workflows, NGS continues to enrich our understanding of inherited conditions and enable faster answers.

Part II. What is Next-Generation Sequencing?

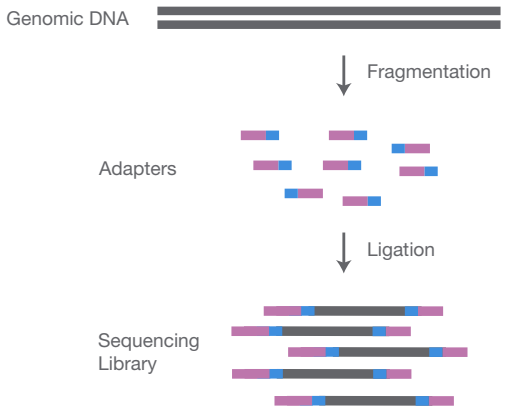
The Basic NGS Workflow

All Illumina NGS workflows include 4 basic steps.

1. Sequencing library preparation begins by creating short DNA or cDNA fragments with 5' and 3' adapters ligated (Figure 1A).
2. For cluster generation, the library is attached to an oligonucleotide lawn on the surface of a flow cell. Through bridge amplification, each library fragment acts as a seed to generate a clonal cluster containing thousands of identical fragments. Across the entire flow cell, millions to billions of clusters are formed (Figure 1B).
3. Next, the templates are ready for sequencing by synthesis (SBS). SBS technology utilizes a proprietary reversible terminator–based method that detects single bases as they are incorporated into DNA template strands.[2] Because all 4 reversible, terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies.[3,4] The result is highly accurate base-by-base sequencing that virtually eliminates sequence-context-specific errors, even within repetitive regions and homopolymers (Figure 1C).
4. The newly identified sequence reads are then exported to an output file and aligned to a reference genome by sequencing alignment software (Figure 1D).

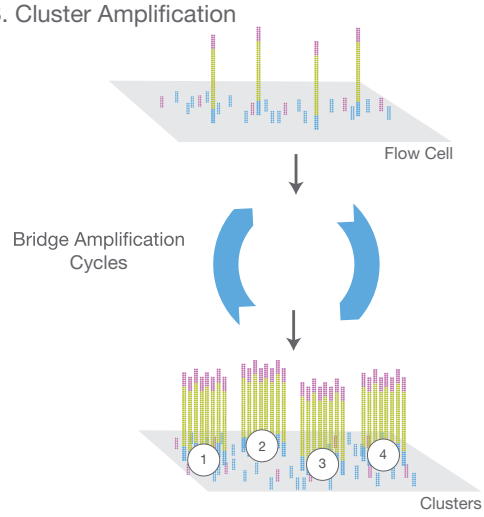
A detailed animation of SBS technology is available at
www.youtube.com/watch?v=HMyCqWhwB8E&list=UUxWMU29FF4kIG8YmQf6Zv0g

A. Library Preparation



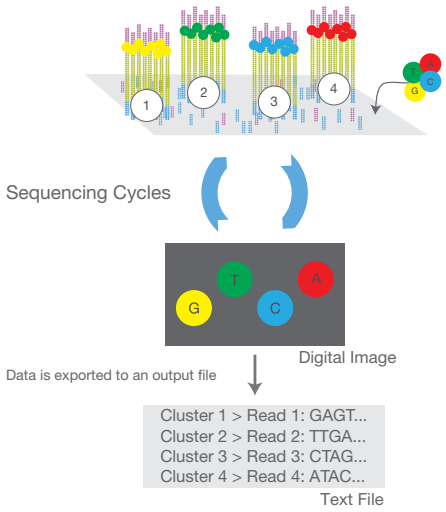
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification



Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

C. Sequencing



Sequencing reagents, including fluorescently labeled nucleotides, are added to the flow cell and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

D. Alignment & Data Anaylsis

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Figure 1: Next-Generation Sequencing (NGS) Workflow—The Illumina NGS workflow follows 4 basic steps: library preparation, cluster generation, sequencing, and data alignment.

Multiplexing

In addition to the rise of data output per run, the sample throughput per run in NGS has also increased over time. Multiplexing allows large numbers or batches of libraries to be pooled and sequenced simultaneously during a single sequencing run (Figure 2). With multiplexing, unique index sequences are added to each DNA fragment during library preparation so that each read can be identified and sorted before final data analysis. This dramatically reduces the time-to-data for multisample studies and enables researchers to go from experiment to answer faster and easier than ever before.

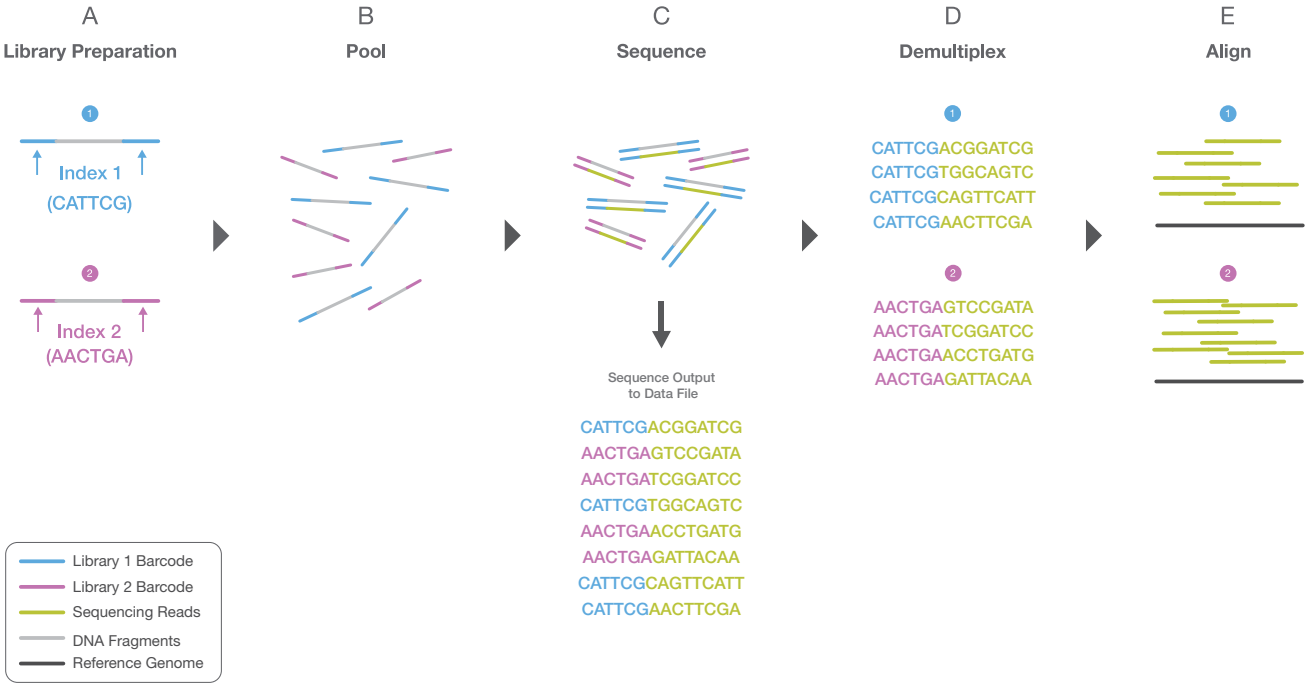


Figure 2: Library Multiplexing Overview—A. Two distinct samples are attached to unique index sequences. Index sequences are attached during library preparation. B. Libraries for each sample are pooled together and loaded into the same flow cell lane. C. Samples are sequenced together during a single instrument run. All sequences are exported to a single output file. D. A demultiplexing algorithm sorts the reads into different files according to their indexes. E. Each set of reads is aligned to the appropriate reference sequence.

Part III. NGS Applications

Next-generation sequencing platforms enable a wide variety of applications, allowing researchers to investigate the genome, transcriptome, or epigenome of any organism. After sequencing libraries are prepared, the actual sequencing process remains fundamentally the same regardless of application.

Whole-Genome Sequencing

Whole-genome sequencing (WGS) involves sequencing the full genome of a given organism. WGS can provide insight into the intronic regions—regions that do not encode proteins—of the genome that exome sequencing does not capture.

Exome Sequencing

The exome—or the protein-coding portion of the human genome—represents less than 2% of the genetic code, but contains ~85% of known disease-related variants. With exome sequencing, the exonic regions of the genome are isolated and sequenced allowing researchers to focus on genetic regions most likely to harbor disease-causing mutations and providing the sensitivity to detect rare variants.

Targeted Sequencing

With targeted sequencing, a pre-selected subset of genes is isolated and sequenced. These customized content sets are referred to as sequencing panels. Due to the tightly defined content, sequencing panels allow the genes within each sample to be sequenced to a greater depth than large-scale approaches like WGS or exome sequencing. This multigene approach confers several advantages:

- Decreases time-to-answer and reduces expenses associated with ordering multiple single-gene tests
- Minimizes issues with limited material for sequential testing due to advances in biopsy procedures
- Enables the accurate identification of rare variants

Part IV. Summary

Over the last decade, advances in NGS technology have led to an improved understanding of genomics, which, in turn, has led to new approaches to managing cardiac disease. Many disease-associated genes and epigenetic markers have been discovered, and ongoing research continues to uncover more.[3-6]

Illumina is committed to providing the highest-quality data in the industry, exemplified by implementation of the largest instrument install base of any NGS technology company[7] and relationships with leaders in the cardiology field. Together, we are bringing the promise of NGS to applications in cardiology, which will lead to widespread clinical adoption, improving patient diagnosis, treatment, and outcomes.

Glossary

Adapters: Specialized oligos bound to the 5' and 3' end of each DNA fragment in a sequencing library. The adapter sequences are complementary to the oligos bound to the surface of Illumina sequencing flow cells.

Bridge amplification: An amplification reaction that occurs on the surface of an Illumina flow cell—also known as "cluster generation." The flow cell surface is coated with a lawn of 2 distinct oligonucleotides. Repeated denaturation and extension cycles (similar to PCR) results in localized amplification of a single fragment into thousands of identical fragments, producing millions to billions of unique, clonal clusters across the flow cell.

Clusters: A clonal grouping of template DNA bound to the surface of a flow cell. Each cluster is seeded by a single, template DNA strand and is clonally amplified through bridge amplification until the cluster has roughly 1000 copies. Each cluster on the flow cell produces a single sequencing read. For example, 1 million clusters on a flow cell would produce 1 million reads.

Flow cell: A glass slide with 1, 2, or 8 (depending on instrument platform) physically separated lanes. Each lane is coated with a lawn of surface bound, adapter-complimentary oligos. A single sample or pool of up to 96 multiplexed samples can be run per lane depending on application parameters.

Indexes: Also known as "barcodes" or "tags"—these are unique sequences, usually 8–12 base pairs long that are ligated to fragments in a sequencing library for identification in subsequent data analysis steps. The index sequences (typically part of the adapter) are added during the library preparation stage.

Multiplexing: Multiple samples, each with a unique index, can be pooled together, loaded into the same flow cell, and sequenced simultaneously during a single sequencing run. Depending on the application and the sequencing instrument used, 10–384 samples can be pooled together.

Read: A unique sequence resulting from a single cluster on the flow cell. The length of the sequence read depends on the number of programmed sequencing cycles during the instrument run. For example, a 150 cycle sequencing run would produce a 150 base pair read. 1 million clusters on the flow cell would result in 1 million unique reads. All sequence reads are exported to a data file following the completion of a sequencing run.

Reference genome: A known, or previously sequenced genome. The reference genome acts as a scaffold against which new sequence reads are aligned (resequencing). In the absence of a reference genome, the genome must be constructed by contig assembly (*de novo* sequencing).

Sequencing by synthesis (SBS): SBS technology uses 4 fluorescently labeled nucleotides to sequence the millions to billions of clusters on a flow cell surface in parallel. During each sequencing cycle, a single labeled dNTP is added to the nucleic acid chain. The nucleotide label serves as a "reversible-terminator" for polymerization. After dNTP incorporation, the fluorescent dye is identified through laser excitation and imaging, then enzymatically cleaved to allow the next round of incorporation. Base calls are made directly from signal intensity measurements during each cycle.[7]

References

1. Ackerman MJ, Priori SG, Willems S, et al. HRS/EHRA expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies. *Europace.* 2011;13(8):1077-1109.
2. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008;456(7218):53-59.
3. Ross MG, Russ C, Costello M, et al. Characterizing and measuring bias in sequence data. *Gen Biol.* 2013;14:R51.
4. CARDIoGRAMplusC4D Consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet.* 2012;45(1):25-33.
5. Ackerman MJ, Priori SG, Willems S, et al. HRS/EHRA expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies: this document was developed as a partnership between the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA). *Europace*. 2011;13(8):1077-1109.
6. Millat G, Bouvagnet P, Chevalier P, et al. Clinical and mutational spectrum in a cohort of 105 unrelated patients with dilated cardiomyopathy. *Eur J Med Genet*. 2011;54(6):e570-575.
7. Nakazato T, Ohta T, Bono H Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*. 2013;8(10):e77910.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY.

Pub. No.770-2014-032 Current as of 05 February 2015

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTA
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA