An introduction to Next-Generation Sequencing Technology

Table of Contents

I. Welcome to Next-Generation Sequencing

a. The Evolution of Genomic Science

DNA sequencing has come a long way since the days of two-dimensional chromatography in the 1970s. With the advent of the Sanger chain termination method[1] in 1977, scientists gained the ability to sequence DNA in a reliable, reproducible manner. A decade later, Applied Biosystems introduced the first automated, capillary electrophoresis (CE)-based sequencing instruments, the AB370 in 1987 and the AB3730xl in 1998, instruments that became the primary workhorses for the NIH-led and Celera-led Human Genome Projects.[2] While these "first-generation" instruments were considered high throughput for their time, the Genome Analyzer emerged in 2005 and took sequencing runs from 84 kilobase (kb) per run to 1 gigabase (Gb) per run.[3] The short read, massively parallel sequencing technique was a fundamentally different approach that revolutionized sequencing capabilities and launched the "next generation" in genomic science. From that point forward, the data output of next-generation sequencing (NGS) has outpaced Moore's law, more than doubling each year (Figure 1).
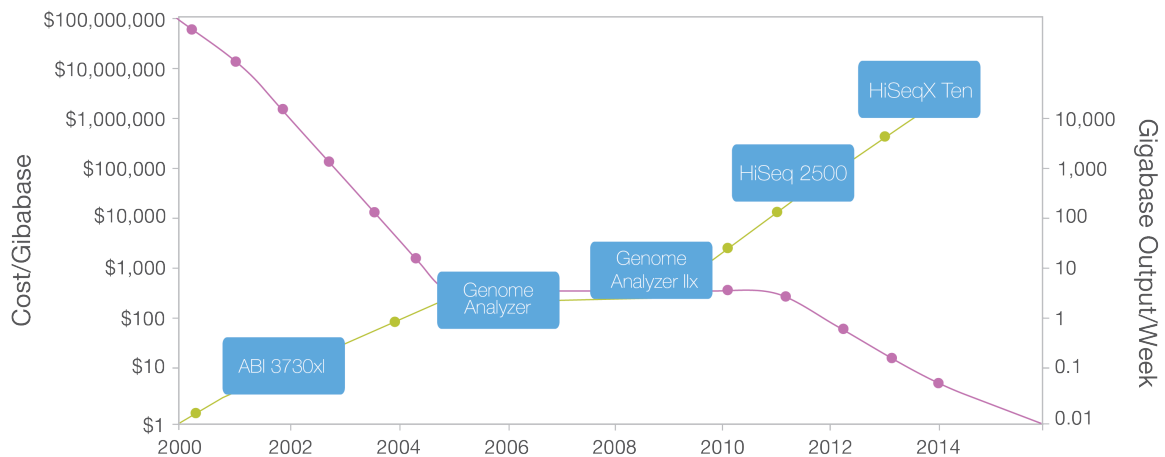

Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

In 2005, a single run on the Genome Analyzer could produce roughly one gigabase of data. By 2014, the rate climbed to 1.8 terabases (Tb) of data in a single sequencing run, an astounding 1000× increase. It is remarkable to reflect on the fact that the first human genome, famously copublished in Science and Nature in 2001, required 15 years to sequence and cost nearly three billion dollars. In contrast, the HiSeq X® Ten System, released in 2014, can sequence over 45 human genomes in a single day for approximately $1000 each (Figure 2).[4]

Beyond the massive increase in data output, the introduction of NGS technology has transformed the way scientists think about genetic information. The $1000 dollar genome enables population-scale sequencing and establishes the foundation for personalized genomic medicine as part of standard medical care. Researchers can now analyze thousands to tens of thousands of samples in a single year. As Eric Lander, founding director of the Broad Institute of MIT and Harvard and principal leader of the Human Genome Project, states:

> "The rate of progress is stunning. As costs continue to come down, we are entering a period where we are going to be able to get the complete catalog of disease genes. This will allow us to look at thousands of people and see the differences among them, to discover critical genes that cause cancer, autism, heart disease, or schizophrenia."[5]

Human Genomes Sequenced Annually

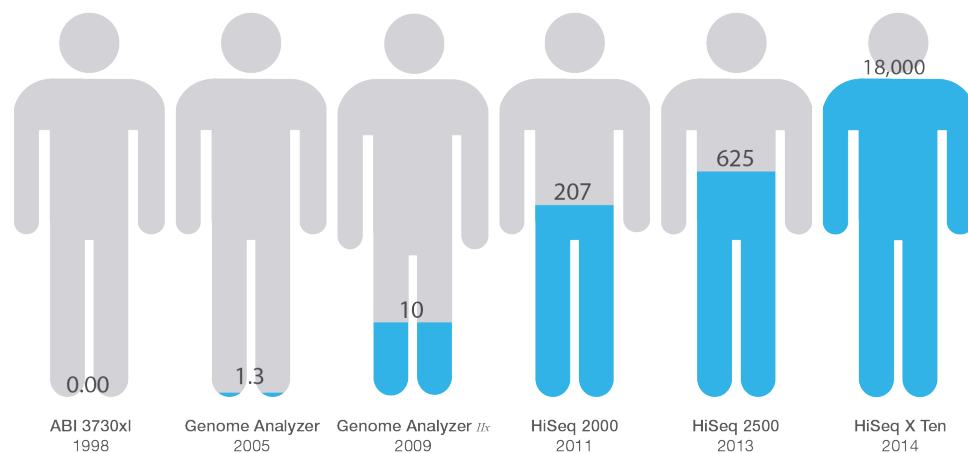| ABI 3730xl	Genome Analyzer	Genome Analyzer *IIx*	HiSeq 2000	HiSeq 2500	HiSeq X Ten
1998	2005	2009	2011	2013	2014
0.00	1.3	10	207	625	18,000

Figure 2: Human Genome Sequencing Over the Decades—The capacity to sequence all 3.2 billion bases of the human genome (at 30× coverage) has increased exponentially since the 1990s. In 2005, with the introduction of the Illumina Genome Analyzer System, 1.3 human genomes could be sequenced annually. Nearly 10 years later, with the Illumina HiSeq X Ten fleet of sequencing systems, the number has climbed to 18,000 human genomes a year.

b. The Basics of NGS Chemistry

In principle, the concept behind NGS technology is similar to CE sequencing. DNA polymerase catalyzes the incorporation of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) into a DNA template strand during sequential cycles of DNA synthesis. During each cycle, at the point of incorporation, the nucleotides are identified by fluorophore excitation. The critical difference is that, instead of sequencing a single DNA fragment, NGS extends this process across millions of fragments in a massively parallel fashion. More than 90% of the world's sequencing data are generated by Illumina sequencing by synthesis (SBS) chemistry.* It delivers high accuracy, a high yield of error-free reads, and a high percentage of base calls above Q30.[6–8]

Illumina NGS workflows include four basic steps:

1. **Library Preparation**—The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5′ and 3′ adapter ligation (Figure 3A). Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process.[9] Adapter-ligated fragments are then PCR amplified and gel purified.

2. **Cluster Generation**—For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification (Figure 3B). When cluster generation is complete, the templates are ready for sequencing.

3. **Sequencing**—Illumina SBS technology uses a proprietary reversible terminator–based method that detects single bases as they are incorporated into DNA template strands (Figure 3C). As all four 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.[6,7] The result is highly accurate base-by-base sequencing that virtually eliminates sequence context–specific errors, even within repetitive sequence regions and homopolymers.

4. **Data Analysis**—During data analysis and alignment, the newly identified sequence reads are aligned to a reference genome (Figure 3D). Following alignment, many variations of analysis are possible, such as single nucleotide polymorphism (SNP) or insertion-deletion (indel) identification, read counting for RNA methods, phylogenetic or metagenomic analysis, and more.

▶ A detailed animation of SBS chemistry is available at www.illumina.com/SBSvideo.

*Data calculations on file. Illumina, Inc., 2015.

A. Library Preparation

Genomic DNA

↓ Fragmentation

Adapters

↓ Ligation

Sequencing Library

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

B. Cluster Amplification

Flow Cell

Bridge Amplification Cycles

1 2 3 4

Clusters

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

C. Sequencing

1 2 3 4

Sequencing Cycles

G T C R

Digital Image

Data is exported to an output file ↓

Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...

Text File

Sequencing reagents, including fluorescently labeled nucleotides, are added 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 and Data Anaylsis

Reads

ATGGCATTGCAATTTGACAT
TGGCATTGCAATTTG
AGATGGTATTG
GATGGCATTGCAA
GCATTGCAATTTGAC
ATGGCATTGCAATT
AGATGGCATTGCAATTTG

Reference Genome

AGATGGTATTGCAATTTGACAT

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.
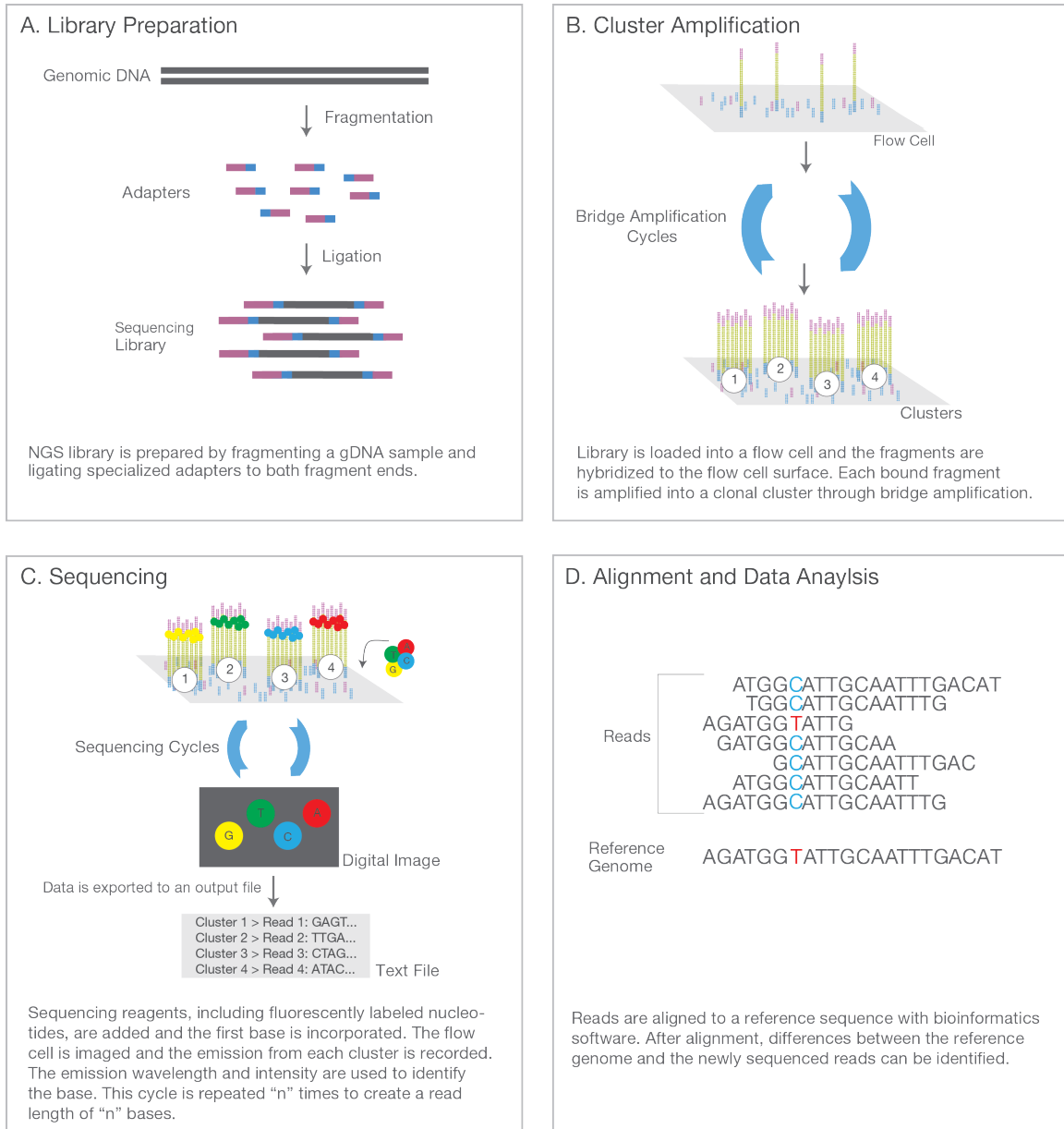
Figure 3: Next-Generation Sequencing Chemistry Overview—Illumina NGS includes four steps: (A) library preparation, (B) cluster generation, (C) sequencing, and (D) alignment and data analysis.

c. Advances in Sequencing Technology

Paired-End Sequencing

A major advance in NGS technology occurred with the development of paired-end (PE) sequencing (Figure 4). PE sequencing involves sequencing both ends of the DNA fragments in a library and aligning the forward and reverse reads as read pairs. In addition to producing twice the number of reads for the same time and effort in library preparation, sequences aligned as read pairs enable more accurate read alignment and the ability to detect indels, which is not possible with single-read data.[8] Analysis of differential read-pair spacing also allows removal of PCR duplicates, a common artifact resulting from PCR amplification during library preparation. Furthermore, PE sequencing produces a higher number of SNV calls following read-pair alignment.[8,9] While some methods are best served by single-read sequencing, such as small RNA sequencing, most researchers currently use the paired-end approach.
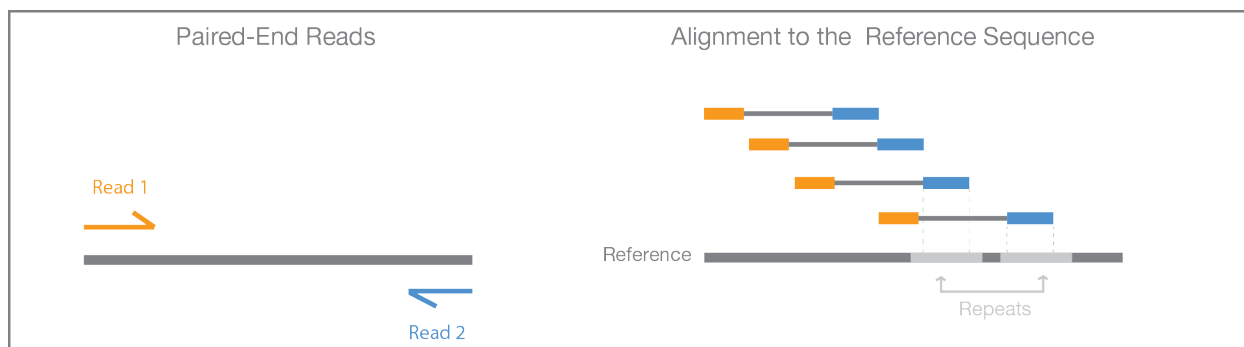
Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better alignment of reads, especially across difficult-to-sequence, repetitive regions of the genome.

Tunable Coverage and Unlimited Dynamic Range

The digital nature of NGS allows a virtually unlimited dynamic range for read-counting methods, such as gene expression analysis. Microarrays measure continuous signal intensities and the detection range is limited by noise at the low end and signal saturation at the high end, while NGS quantifies discrete, digital sequencing read counts. By increasing or decreasing the number of sequencing reads, researchers can tune the sensitivity of an experiment to accommodate various study objectives. Because the dynamic range with NGS is adjustable and nearly unlimited, researchers can quantify subtle gene expression changes with much greater sensitivity than traditional microarray-based methods. Sequencing runs can be tailored to zoom in with high resolution on particular regions of the genome, or provide a more expansive view with lower resolution.

The ability to easily tune the level of coverage offers several experimental design advantages. For instance, somatic mutations may only exist within a small proportion of cells in a given tissue sample. Using mixed tumor–normal cell samples, the region of DNA harboring the mutation must be sequenced at extremely high coverage, often upwards of 1000×, to detect these low-frequency mutations within the mixed cell population. On the other side of the coverage spectrum, a method like genome-wide variant discovery usually requires a much lower coverage level. In this case, the study design involves sequencing many samples (hundreds to thousands) at lower resolution, to achieve greater statistical power within a given population.

Advances in Library Preparation

With Illumina NGS, library preparation has undergone rapid improvements. The first NGS library prep protocols involved random fragmentation of the DNA or RNA sample, gel-based size selection, ligation of platform-specific oligonucleotides, PCR amplification, and several purification steps. While the 1–2 days required to generate these early NGS libraries were a great improvement over traditional cloning techniques, current NGS protocols, such as Nextera® XT DNA Library Preparation, have reduced the library prep time to less than 90 minutes.[10] PCR-free and gel-free kits are also available for sensitive sequencing methods. PCR-free library preparation kits result in superior coverage of traditionally challenging areas such as high AT/GC-rich regions, promoters, and homopolymeric regions.[11]

🔗 For a complete list of Illumina library preparation kits, visit www.illumina.com/products/by-type/sequencing-kits/library-prep-kits.html.

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 of libraries to be pooled and sequenced simultaneously during a single sequencing run (Figure 5). With multiplexed libraries, 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. With PE sequencing and multiplexing, NGS has dramatically reduced the time to data for multisample studies and enabled researchers to go from experiment to data quickly and easily.

Gains in throughput from multiplexing come with an added layer of complexity, as sequencing reads from pooled libraries need to be identified and sorted computationally in a process called demultiplexing before final data analysis (Figure 5). The phenomenon of index misassignment between multiplexed libraries is a known issue that has impacted NGS technologies from the time sample multiplexing was developed.[12] Index hopping is a specific cause of index misassignment that can result in incorrect assignment of libraries from the expected index to a different index in the pool, leading to misalignment and inaccurate sequencing results.

For more information regarding index hopping, including mechanisms by which it occurs, how Illumina measures index hopping, and best practices for mitigating the impact of index hopping on sequencing data quality, read the Effects of Index Misassignment on Multiplexing and Downstream Analysis White Paper.
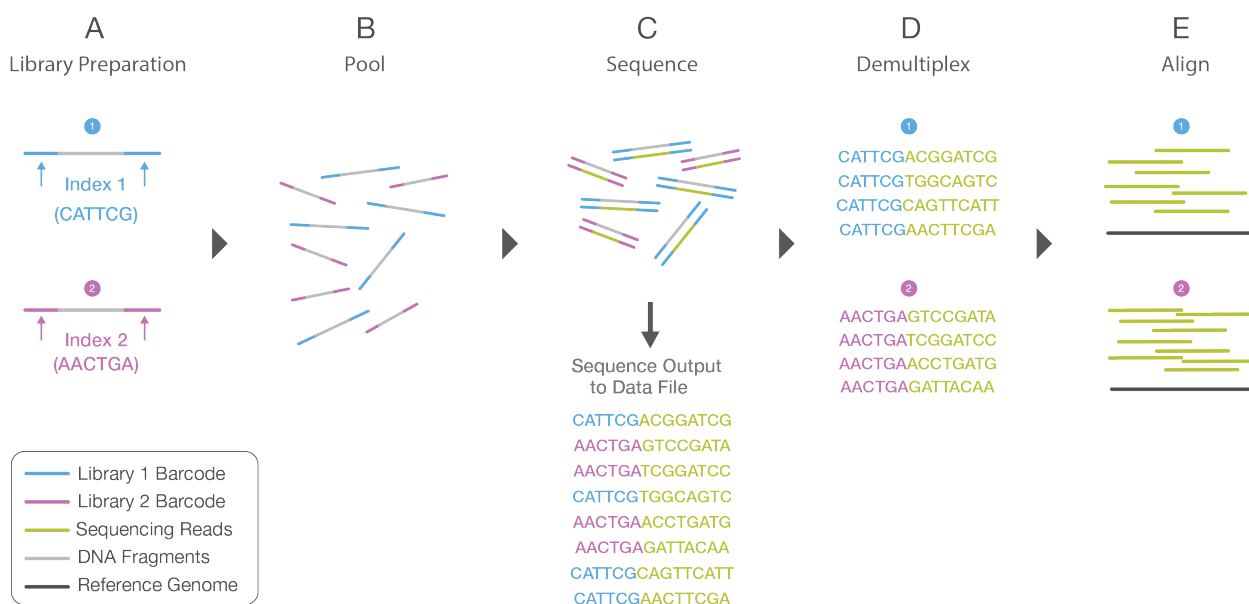

Figure 5: Library Multiplexing Overview—(A) Unique index sequences are added to two different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single instrument run. All sequences are exported 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.

Flexible, Scalable Instrumentation

While the latest NGS platforms can produce massive data output, NGS technology is also highly flexible and scalable. Sequencing systems are available for every method and scale of study, from small laboratories to large genome centers (Figure 6). Illumina NGS instruments range from the benchtop MiniSeq™ System, with output ranging from 1.8–7.5 Gb for targeted sequencing studies, to the NovaSeq™ 6000 System, which can generate an impressive 6 Tb and 20 B reads in ~ 2 days[†] for population-scale studies.

Flexible run configurations are also engineered into the design of Illumina NGS sequencers. For example, the HiSeq® 2500 System offers two run modes and single or dual flow cell sequencing while the NextSeq® Series of Sequencing Systems offers two flow cell types to accommodate different throughput requirements. The HiSeq 3000/4000 Series uses the same patterned flow cell technology as the HiSeq X instruments for cost-effective production-scale sequencing. The new NovaSeq Series of systems unites the latest high-performance imaging with the next generation of Illumina patterned flow cell

technology to deliver massive increases in throughput. This flexibility allows researchers to configure runs tailored to their specific study requirements, with the instrument of their choice.

For an in-depth comparison of Illumina platforms, visit www.illumina.com/systems/sequencing.html or explore the Sequencing Platform Comparison Tool at www.illumina.com/systems/sequencing-platforms/comparison-tool.html.

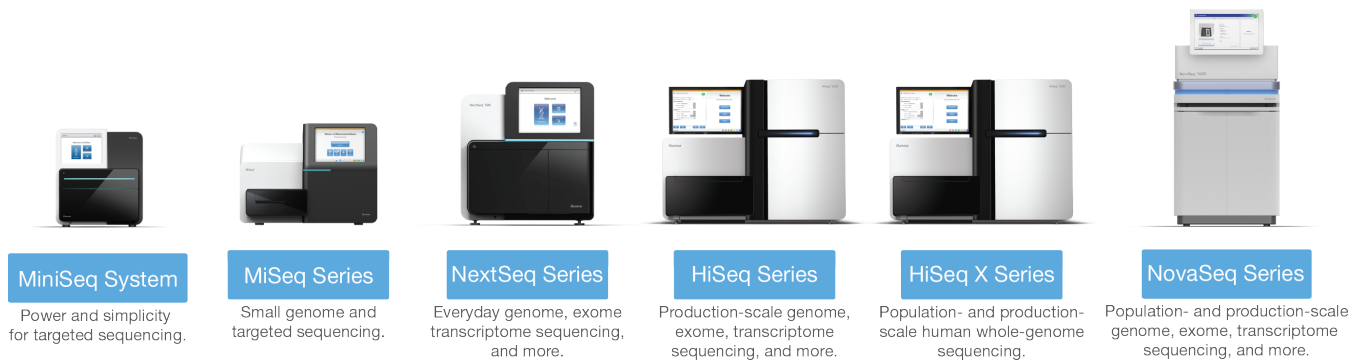| MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series	NovaSeq Series
Power and simplicity for targeted sequencing.	Small genome and targeted sequencing.	Everyday genome, exome transcriptome sequencing, and more.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale human whole-genome sequencing.	Population- and production-scale genome, exome, transcriptome sequencing, and more.

Figure 6: Sequencing Systems for Virtually Every Scale—Illumina offers innovative NGS platforms that deliver exceptional data quality and accuracy over a wide scale, from small benchtop sequencers to production-scale sequencing systems.

II. NGS Methods

NGS platforms enable a wide variety of methods, allowing researchers to ask virtually any question related to the genome, transcriptome, or epigenome of any organism. Sequencing methods differ primarily by how the DNA or RNA samples are obtained (eg, organism, tissue type, normal vs. affected, experimental conditions, etc) and by the data analysis options used. After the sequencing libraries are prepared, the actual sequencing stage remains fundamentally the same, regardless of the method. There are various standard library preparation kits that offer protocols for whole-genome sequencing (WGS), RNA sequencing (RNA-Seq), targeted sequencing (such as exome sequencing or 16S sequencing), custom-selected regions, protein-binding regions, and more. Although the number of NGS methods is constantly growing, a brief overview of the most common methods is presented here.

a. Genomics

Whole-Genome Sequencing

Microarray-based, genome-wide association studies (GWAS) have been a common approach for identifying disease associations across the whole genome. While GWAS microarrays can interrogate over four million markers per sample, the most comprehensive method of interrogating the 3.2 billion bases of the human genome is WGS. The rapid drop in sequencing cost and the ability of WGS to produce large volumes of data rapidly make it a powerful tool for genomics research. While WGS is commonly associated with sequencing human genomes, the scalable, flexible nature of the method makes it equally useful for sequencing any species, such as agriculturally important livestock, plant genomes, or disease-related microbial genomes. This broad utility was demonstrated during the recent *E. coli* outbreak in Europe in 2011, which prompted a rapid scientific response. Using the latest NGS systems, researchers quickly sequenced the bacterial strain, enabling them to track the origins and transmission of the outbreak as well as identify genetic mutations conferring the increased virulence.[13]

Exome Sequencing

Exome sequencing is a widely-used targeted sequencing method. The exome represents less than 2% of the human genome, but contains most of the known disease-causing variants, making whole-exome sequencing (WES) a cost-effective alternative to WGS.[14] With WES, the protein-coding portion of the genome is selectively captured and sequenced. It can efficiently identify variants across a wide range of applications, including population genetics, genetic disease, and cancer studies.

---

†With dual flow cell mode enabled.

De novo Sequencing

De novo sequencing refers to sequencing a novel genome where there is no reference sequence available for alignment. Sequence reads are assembled as contigs and the coverage quality of *de novo* sequence data depends on the size and continuity of the contigs (ie, the number of gaps in the data). Another important factor in generating high-quality *de novo* sequences is the diversity of insert sizes included in the library. Combining short-insert paired-end and long-insert mate pair sequences is the most powerful approach for maximal coverage across the genome (Figure 7). The combination of insert sizes enables detection of the widest range of structural variant types and is essential for accurately identifying more complex rearrangements. The short-insert reads, sequenced at higher depths, can fill in gaps not covered by the long inserts, which are often sequenced at lower read depths. Therefore, using a combined approach results in higher quality assemblies. In parallel with NGS technology improvements, many algorithmic advances have emerged in sequence assemblers for short-read data. Researchers can perform high-quality *de novo* assembly using NGS reads and publicly available short-read assembly tools with existing computer resources in the laboratory.
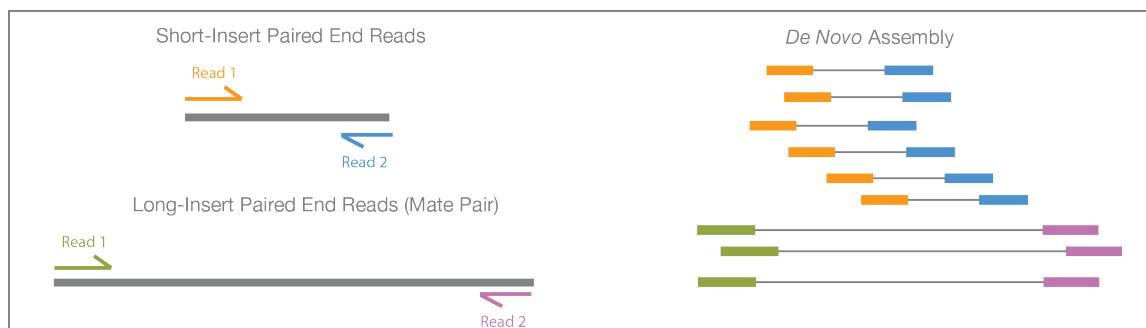

Figure 7: Mate Pairs and *De novo* Assembly—Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for *de novo* assembly.

Targeted Sequencing

With targeted sequencing, a subset of genes or regions of the genome are isolated and sequenced. Targeted sequencing allows researchers to focus time, expenses, and data analysis on specific areas of interest and enables sequencing at much higher coverage levels. For example, a typical WGS study achieves coverage levels of 30–50× per genome, while a targeted resequencing project can easily cover the target region at 500–1000× or higher. This higher coverage allows researchers to identify rare variants, variants that would be too rare and too expensive to identify with WGS or CE-based sequencing.

Targeted sequencing panels can be purchased with fixed, preselected content or can be custom designed. A wide variety of targeted sequencing library prep kits are available, including kits with probe sets focused on specific areas of interest such as cancer, cardiomyopathy, or autism. Custom probe sets are available through DesignStudio™ Software enabling researchers to target regions of the genome relevant to specific research interests. Custom targeted sequencing is ideal for examining genes in specific pathways, or for follow-up studies from GWAS or WGS. Illumina currently supports two methods for targeted sequencing, target enrichment and amplicon generation (Figure 8).

Target enrichment captures between 10 kb–62 Mb regions, depending on the library prep kit parameters. Amplicon sequencing allows researchers to sequence 16–1536 targets at a time, spanning 2.4–652.8 kb of total content, depending on the library prep kit used. This highly multiplexed approach enables a wide range of applications for discovery, validation, or screening of genetic variants. Amplicon sequencing is useful for discovery of rare somatic mutations in complex samples (eg, cancerous tumors mixed with germline DNA).[15,16] Another common amplicon application is sequencing the bacterial 16S rRNA gene across multiple species, a widely used method for phylogeny and taxonomy studies, particularly in diverse metagenomic samples.[17]

For more information on Illumina targeted, WGS, exome, or *de novo* sequencing solutions, visit www.illumina.com/applications/sequencing/dna_sequencing.html.
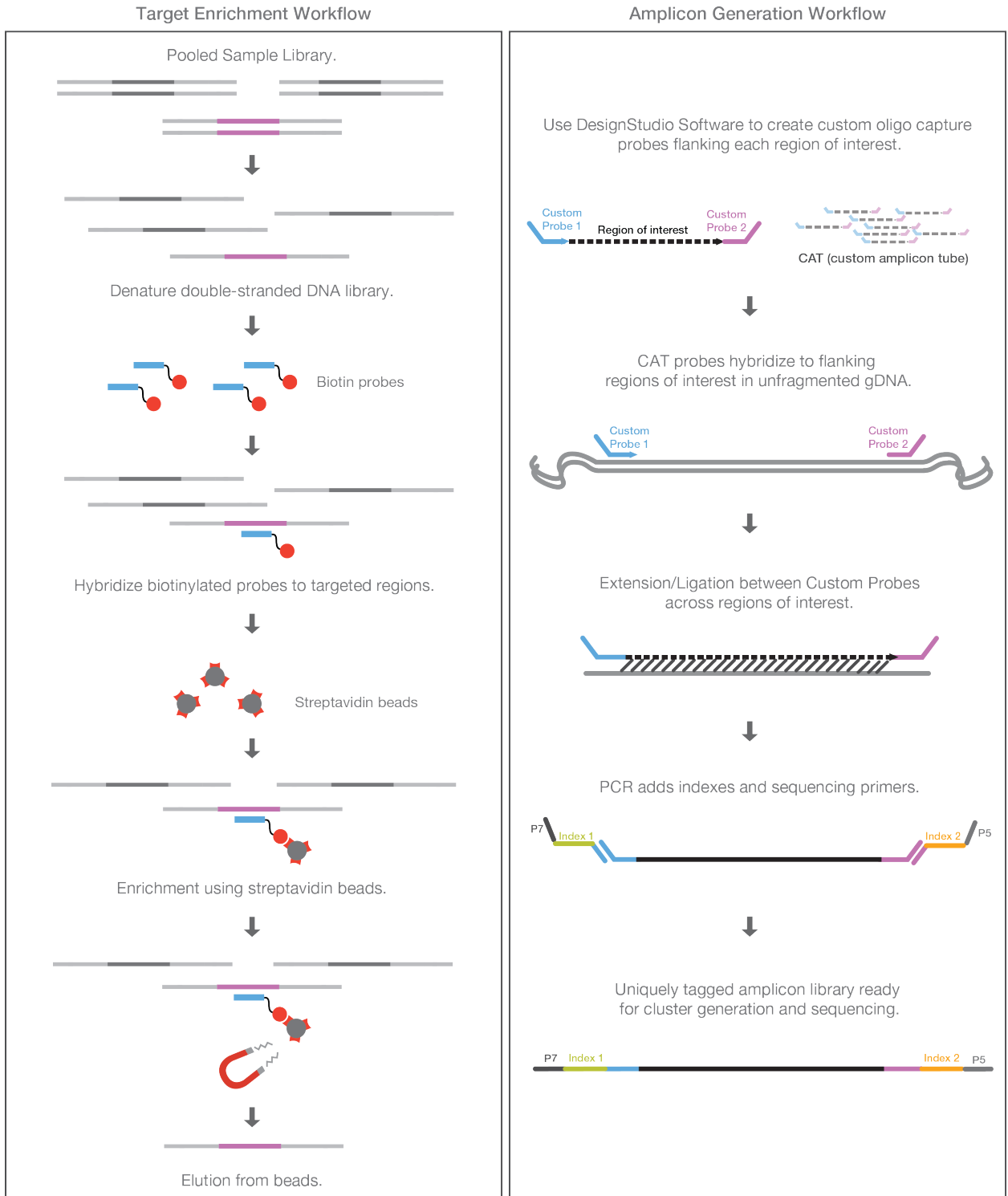
Target Enrichment Workflow

Pooled Sample Library.

Denature double-stranded DNA library.

Biotin probes

Hybridize biotinylated probes to targeted regions.

Streptavidin beads

Enrichment using streptavidin beads.

Elution from beads.

Amplicon Generation Workflow

Use DesignStudio Software to create custom oligo capture probes flanking each region of interest.

Custom Probe 1   Region of interest   Custom Probe 2

CAT (custom amplicon tube)

CAT probes hybridize to flanking regions of interest in unfragmented gDNA.

Custom Probe 1   Custom Probe 2

Extension/Ligation between Custom Probes across regions of interest.

PCR adds indexes and sequencing primers.

P7   Index 1   Index 2   P5

Uniquely tagged amplicon library ready for cluster generation and sequencing.

P7   Index 1   Index 2   P5

Figure 8: Target Enrichment and Amplicon Generation Workflows—With target enrichment, specific regions of interest are captured by hybridization to biotinylated probes, then isolated by magnetic pulldown. Amplicon sequencing involves the amplification and purification of regions of interest using highly multiplexed PCR oligos sets.

b. Transcriptomics

Library preparation methods for RNA-Seq typically begin with total RNA sample preparation followed by a ribosome removal step. The total RNA sample is then converted to cDNA before standard NGS library preparation. RNA-Seq focused on mRNA, small RNA, noncoding RNA, or microRNAs can be achieved by including additional isolation or enrichment steps before cDNA synthesis (Figure 9).
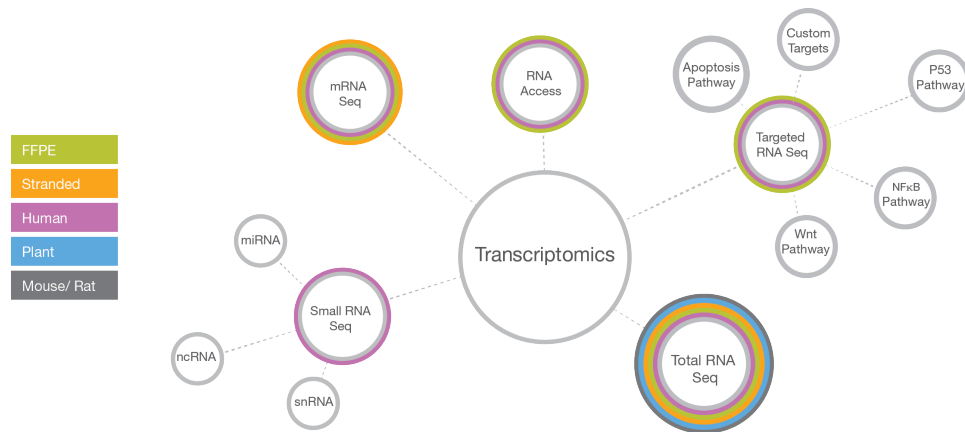

Figure 9: A Complete View of Transcriptomics with NGS—A broad range of methods for transcriptomics with NGS have emerged over the past 10 years including total RNA-Seq, mRNA-Seq, small RNA-Seq, and targeted RNA-Seq.

Total RNA and mRNA Sequencing

Transcriptome sequencing is a major advance in the study of gene expression because it allows a snapshot of the whole transcriptome rather than a predetermined subset of genes. Whole-transcriptome sequencing provides a comprehensive view of a cellular transcriptional profile at a given biological moment and greatly enhances the power of RNA discovery methods. As with any sequencing method, an almost unlimited dynamic range allows identification and quantification of both common and rare transcripts. Additional capabilities include aligning sequencing reads across splice junctions, and detection of isoforms, novel transcripts, and gene fusions. Library preparation kits that support precise detection of strand orientation are available for both total RNA-Seq and mRNA-Seq methods.

Targeted RNA Sequencing

Targeted RNA sequencing is a method for measuring transcripts of interest for detecting differential expression, allele-specific expression, detection of gene-fusions, isoforms, cSNPs, and splice junctions. Illumina TruSeq® Targeted RNA Sequencing Kits include preconfigured, experimentally validated panels focused on specific cellular pathways or disease states such as apoptosis, cardiotoxicity, NFκB pathway, and more. Custom content can be designed and ordered for analysis of specific genes of interest. Targeted RNA sequencing is a powerful method for the investigation of specific pathways of interest or for the validation of gene expression microarray or whole-transcriptome sequencing results.

Small RNA and Noncoding RNA Sequencing

Small, noncoding RNA, or microRNAs are short, 18–22 bp nucleotides that play a role in the regulation of gene expression often as gene repressors or silencers. The study of microRNAs has grown as their role in transcriptional and translational regulation has become more evident.[18,19]

For more information regarding Illumina solutions for small RNA (noncoding RNA), targeted RNA, total RNA, and mRNA sequencing, visit www.illumina.com/applications/sequencing/rna.html.

c. Epigenomics

While genomics involves the study of heritable or acquired alterations in the DNA sequence, epigenetics is the study of heritable changes in gene activity caused by mechanisms other than DNA sequence changes. Mechanisms of epigenetic activity include DNA methylation, small RNA–mediated regulation, DNA–protein interactions, histone modification, and more.

Methylation Sequencing

A critical focus in epigenetics is the study of cytosine methylation (5mC) states across specific areas of regulation, such as promotors or heterochromatin. Cytosine methylation can significantly modify temporal and spatial gene expression and chromatin remodeling.[20] While there are many methods for the study of genetic methylation, methylation sequencing leverages the advantages of NGS technology and genome-wide analysis while assessing methylation states at the single-nucleotide level. Two methylation sequencing methods are widely used: whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). With WGBS, sodium bisulfite chemistry converts nonmethylated cytosines to uracils, which are then converted to thymines in the sequence reads or data output. In RRBS, DNA is digested with MspI, a restriction enzyme unaffected by methylation status. Fragments in the 100–150 bp size range are isolated to enrich for CpG and promotor containing DNA regions. Sequencing libraries are then constructed using the standard NGS protocols.

> For more information on methylation sequencing solutions, visit
> www.illumina.com/techniques/sequencing/methylation-sequencing.html

ChIP Sequencing

Protein–DNA or protein–RNA interactions have a significant impact on many biological processes and disease states. These interactions can be surveyed with NGS by combining chromatin immunoprecipitation (ChIP) assays and NGS methods. ChIP-Seq protocols begin with the chromatin immunoprecipitation step (ChIP protocols vary widely as they must be specific to the species, tissue type, and experimental conditions).

> For more information on ChIP-Seq, visit
> www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html.

Ribosome Profiling

Ribosome profiling is a method based on deep sequencing of ribosome protected–mRNA fragments. Purification and sequencing of these fragments provides a "snapshot" of all the ribosomes active in a cell at a specific time point. This information can determine what proteins are being actively translated in a cell, and can be useful for investigating translational control, measuring gene expression, determining the rate of protein synthesis, or predicting protein abundance. Ribosome profiling enables systematic monitoring of cellular translation processes and prediction of protein abundance. Determining what regions of a transcript are being translated can help define the proteome of complex organisms. With NGS, ribosome profiling allows detailed and accurate *in vivo* analysis of protein production.

> To learn more about Illumina ribosome profiling, visit
> www.illumina.com/applications/sequencing/rna.html.

III. Illumina DNA-to-Data NGS Solutions

a. The Illumina NGS Workflow

Illumina offers a comprehensive solution for the NGS workflow, from library preparation to data analysis (Figure 10). Library preparation kits are available for all NGS methods, including WGS, exome sequencing, targeted sequencing, RNA-Seq, and more. Illumina library preparation protocols can accommodate a range of throughput needs, from manual protocols for smaller laboratories to fully automated library preparation workstations for larger laboratories or genome centers. Likewise, Illumina offers a full portfolio of sequencing platforms, from the benchtop MiniSeq and MiSeq® Systems to the factory-scale HiSeq X and NovaSeq Series of Sequencing Systems that deliver the right level of speed, capacity, and cost for various laboratory or sequencing center requirements. For the last step in the NGS workflow, Illumina offers user-friendly bioinformatics tools that are easily accessible through the web, on instrument, or through onsite servers.

Figure 10: Illumina DNA-to-Data Solutions—Illumina provides fully integrated, DNA-to-data solutions, with technology and support for every step of the NGS workflow including library preparation, sequencing, and final data analysis.

b. Integrated Data Analysis

Data from any Illumina sequencing system can be streamed into BaseSpace® Sequence Hub, a user-friendly genomics cloud computing platform that offers simplified data management, analytical sequencing tools, and data storage. BaseSpace Sequence Hub is optimized to automate processing of the large volume of data generated. Researchers will find a rich ecosystem of commercial and open-source tools, from Illumina and third-party developers, for data analysis, including alignment and variant detection, annotation, visualization, interpretation, and somatic variant calling. BaseSpace Onsite Sequence Hub is a local version of BaseSpace Sequence Hub that enables data storage and analysis onsite through an installed local server. On-instrument access to BaseSpace Sequence Hub enables the integration of many workflow steps, including library prep planning with BaseSpace Prep,[‡] run set-up and chemistry validation, and real-time automatic data transfer to the BaseSpace computing environment.

The NGS workflow then proceeds seamlessly through alignment and subsequent data analysis steps with BaseSpace Apps. BaseSpace Apps offer a wide variety of analysis pipelines, including analysis for *de novo* assembly, SNP and indel variant analysis, RNA expression profiling, 16S metagenomics, tumor-normal comparisons, epigenetic/gene regulation analysis, and many more. Illumina collaborates closely with commercial and academic software developers to create a full ecosystem of data analysis tools that address the needs of various research objectives. In the final stages of the NGS workflow, data can be shared with collaborators or delivered instantly to customers around the world.[§]

To learn more about BaseSpace Sequence Hub, visit
www.illumina.com/basespace.

[‡]Currently available with MiniSeq and NextSeq 500/550 Systems only. HiSeq and MiSeq Systems can use Illumina Experiment Manager (IEM) for the same planning and validation functions.

[§]Cloud-based environment only. BaseSpace Onsite Sequence Hub restricts data sharing to local users.

IV. Glossary

adapters: The oligos bound to the 5' and 3' end of each DNA fragment in a sequencing library. The adapters are complementary to the lawn of oligos present on the surface of Illumina sequencing flow cells.

bridge amplification: An amplification reaction that occurs on the surface of an Illumina flow cell. During flow cell manufacturing, the surface is coated with a lawn of two distinct oligonucleotides often referred to as "P5" and "P7." In the first step of bridge amplification, a single-stranded sequencing library (with complementary adapter ends) is loaded into the flow cell. Individual molecules in the library bind to complementary oligos as they "flow" across the oligo lawn. Priming occurs as the opposite end of a ligated fragment bends over and "bridges" to another complementary oligo on the surface. Repeated denaturation and extension cycles (similar to PCR) results in localized amplification of single molecules into millions of unique, clonal clusters across the flow cell. This process, also known as "clustering," occurs in an automated, flow cell instrument called a cBot System or in an onboard cluster module within an NGS instrument.

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 ~1000 copies. Each cluster on the flow cell produces a single sequencing read. For example, 10,000 clusters on the flow cell would produce 10,000 single reads and 20,000 paired-end reads.

contigs: A stretch of continuous sequence, *in silico*, generated by aligning overlapping sequencing reads.

coverage level: The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was sequenced 30 times.

flow cell: A glass slide with one, two, or eight physically separated lanes, depending on the instrument platform. Each lane is coated with a lawn of surface bound, adapter-complimentary oligos. A single library or a pool of up to 96 multiplexed libraries can be run per lane, depending on application parameters.

indexes/barcodes/tags: A unique DNA sequence ligated to fragments within a sequencing library for downstream *in silico* sorting and identification. Indexes are typically a component of adapters or PCR primers and are ligated to the library fragments during the sequencing library preparation stage. Illumina indexes are typically between 8–12 bp. Libraries with unique indexes can be pooled together, loaded into one lane of a sequencing flow cell, and sequenced in the same run. Reads are later identified and sorted via bioinformatic software. All together, this process is known as "multiplexing."

insert: During the library preparation stage, the sample DNA is fragmented, and the fragments of a specific size (typically 200–500 bp, but can be larger) are ligated or "inserted" in between two oligo adapters. The original sample DNA fragments are also referred to as "inserts."

mate pair library: A sequencing library with long inserts ranging in size from 2–5 kb typically run as paired-end libraries. The long gap length in between the sequence pairs is useful for building contigs in *de novo* sequencing, identification of indels, and other methods.

multiplexing: See "indexes/barcodes/tags."

read: NGS uses sophisticated instruments to determine the nucleotide sequence of a DNA or RNA sample. In general terms, a sequence "read" refers to the data string of A, T, C, and G bases corresponding to the sample DNA or RNA. With Illumina technology, millions of reads are generated in a single sequencing run.

reference genome: A reference genome is a fully sequenced and assembled genome that acts as a scaffold against which new sequence reads are aligned and compared. Typically, reads generated from a sequencing run are aligned to a reference genome as a first step in data analysis.

sequencing by synthesis (SBS): SBS technology uses four fluorescently labeled nucleotides to sequence the tens of millions of clusters on the 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.

V. References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *PNAS*. 1977;74(12):5463–5467.
2. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. *Science*. 2003;300(5617):286–290.
3. Davies K. 13 years ago, a beer summit in an English pub led to the birth of Solexa. *BioIT World*. September 28, 2010. (www.bio-itworld.com/2010/issues/sept-oct/solexa.html).
4. Illumina. HiSeq X Ten Series of Sequencing Systems. 2014. (www.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf).
5. Fallows J. When will genomics cure cancer?. *The Atlantic*. January 2014. (www.theatlantic.com/magazine/archive/2014/01/when-will-genomics-cure-cancer/355739/)
6. Ross MG, Russ C, Costello M, et al. Characterizing and measuring bias in sequence data. *Genome Biol*. 2013;14(5):R51.
7. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nat*. 2008;456(7218):53–59.
8. 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.
9. Illumina. Nextera DNA Library Preparation Kits data sheet. 2014. (www.illumina.com/documents/products/datasheets/datasheet_nextera_dna_sample_prep.pdf).
10. Illumina. Nextera XT DNA Library Preparation Kit data sheet. 2014.(www.illumina.com/documents/products/datasheets/datasheet_nextera_xt_dna_sample_prep.pdf).
11. Illumina. TruSeq DNA PCR-Free Library Preparation Kit data sheet. 2013. (www.illumina.com/documents/products/datasheets/datasheet_truseq_dna_pcr_free_sample_prep.pdf).
12. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. 2012:2513–2524.
13. Grad YH, Lipsitch M, Feldgarden M, et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. *PNAS*. 2012;109(8):3065–3070.
14. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;(9):418–426.
15. McEllistrem MC. Genetic diversity of the pneumococcal capsule: implications for molecular-based serotyping. *Future Microbiol*. 2009;4(7):857–865.
16. Lo YMD, Chiu RWK. Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. *Clin Chem*. 2009;55:607–608.
17. Ram JL, Karim AS, Sendler ED, Kato I. Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the Illumina sequencing platform. *Syst Biol Reprod Med*. 2011;57(3):117–118.
18. Wang Y, Kim S, Kim IM. Regulation of metastasis by microRNAs in ovarian cancer. *Front Oncol*. 2014;10:143.
19. Dior Up, Kogan L, Chill HH, Eizenberg N, Simon A, Revel A. Emerging roles of microRNA in the embryo-endometrium cross talk. *Semin Reprod Med*. 2014;32(5):402–409.
20. Phillips T. The role of methylation in gene expression. *Nat Education*. 2008;1(1):116.

For Research Use Only. Not for use in diagnostic procedures.