An Introduction to Next-Generation Sequencing Technology

Deciphering DNA sequences is essential for virtually all branches of biological research. With the advent of capillary electrophoresis (CE)-based Sanger sequencing, scientists gained the ability to elucidate genetic information from any given biological system. This technology has become widely adopted in laboratories around the world, yet has always been hampered by inherent limitations in throughput, scalability, speed, and resolution that often preclude scientists from obtaining the essential information they need for their course of study. To overcome these barriers, an entirely new technology was required—Next-Generation Sequencing (NGS), a fundamentally different approach to sequencing that triggered numerous ground-breaking discoveries and ignited a revolution in genomic science.

www.illumina.com/NGS

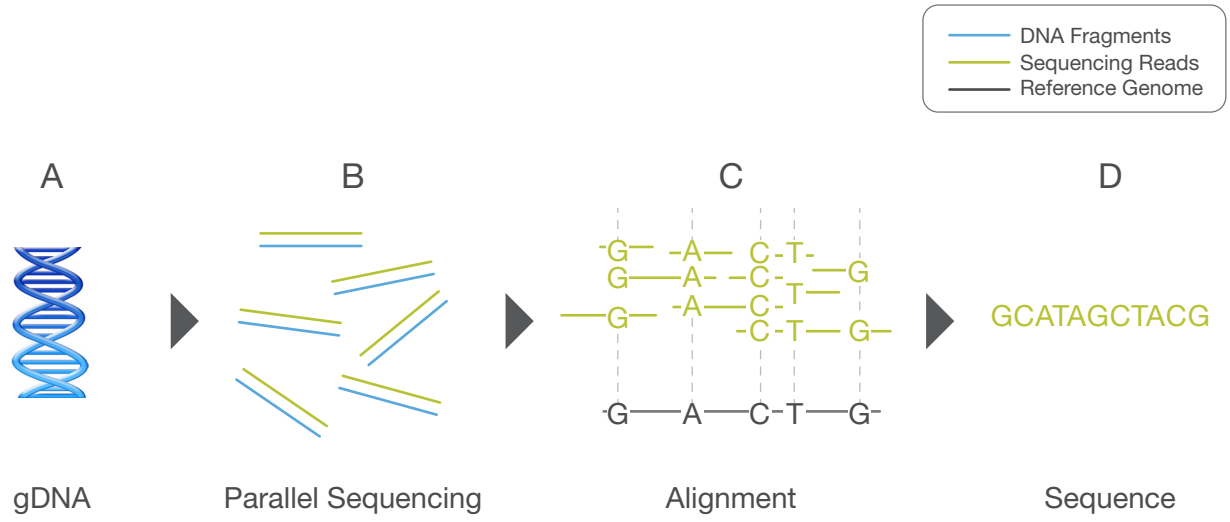
illumina®

Welcome to Next-Generation Sequencing

The five years since the introduction of NGS technology have seen a major transformation in the way scientists extract genetic information from biological systems, revealing limitless insight about the genome, transcriptome, and epigenome of any species. This ability has catalyzed a number of important breakthroughs, advancing scientific fields from human disease research to agriculture and evolutionary science.

In principle, the concept behind NGS technology is similar to CE—the bases of a small fragment of DNA are sequentially identified from signals emitted as each fragment is re-synthesized from a DNA template strand. NGS extends this process across millions of reactions in a massively parallel fashion, rather than being limited to a single or a few DNA fragments. This advance enables rapid sequencing of large stretches of DNA base pairs spanning entire genomes, with the latest instruments capable of producing hundreds of gigabases of data in a single sequencing run. To illustrate how this process works, consider a single genomic DNA (gDNA) sample. The gDNA is first fragmented into a library of small segments that can be uniformly and accurately sequenced in millions of parallel reactions. The newly identified strings of bases, called reads, are then reassembled using a known reference genome as a scaffold (resequencing), or in the absence of a reference genome (*de novo* sequencing). The full set of aligned reads reveals the entire sequence of each chromosome in the gDNA sample (Figure 1).

Figure 1: Conceptual Overview of Whole-Genome Resequencing



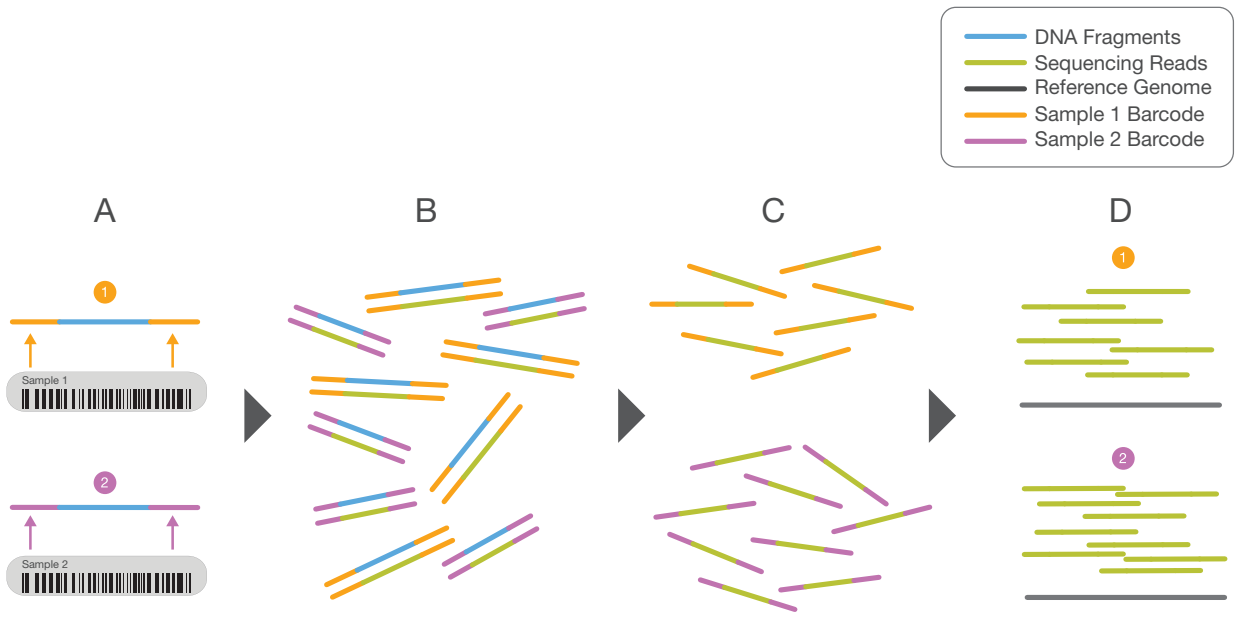
A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

High-Throughput Science

NGS data output has increased at a rate that outpaces Moore's law, more than doubling each year since it was invented. In 2007, a single sequencing run could produce a maximum of around one gigabase (Gb) of data. By 2011, that rate has nearly reached a terabase (Tb) of data in a single sequencing run—nearly a 1000× increase in four years. With the ability to rapidly generate large volumes of sequencing data, NGS enables researchers to move quickly from an idea to full data sets in a matter of hours or days. Researchers can now sequence more than five human genomes in a single run, producing data in roughly one week, for a reagent cost of less than $5,000 per genome. By comparison, the first human genome required roughly 10 years to sequence using CE technology and an additional three years to finish the analysis. The completed project was published in 2003, just a few years before NGS was invented, and came with a price tag nearing 3 billion USD.

While the latest high-throughput sequencing instruments are capable of massive data output, NGS technology is highly scalable. The same underlying chemistry can be used for lower output volumes for targeted studies or smaller genomes. This scalability gives researchers the flexibility to design studies that best suit the needs of their particular research. For sequencing small bacterial/viral genomes or targeted regions like exomes, a researcher can choose to use a lower output instrument and process a smaller number of samples per run, or can opt to process a large number of samples by multiplexing on a high-throughput instrument. Multiplexing enables large sample numbers to be simultaneously sequenced during a single experiment (Figure 2). To accomplish this, individual "barcode" sequences are added to each sample so they can be differentiated during the data analysis.

Figure 2: Conceptual Overview of Sample Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

With multiplexing, NGS dramatically reduces the time to data for multi-sample studies. Processing hundreds of amplicons using CE technology generally requires several weeks or months. The same number of samples can now be sequenced in a matter of hours and fully analyzed within two days using NGS. With highly automated, easy-to-use protocols, researchers can go from experiment to data to publication faster and easier than ever before (Table 1).

Table 1: A comparison of Illumina NGS and CE-Based Sanger Sequencing

Technology	Starting Material	Samples per Run	Run Time*	Read Length	Number of Reads	Output per Run	Applications
CE-based Sanger Method	1–3 µg	1-96	0.5 hrs	550 bp†	1–96	0.550–52.5 kb	DNA sequencing, resequencing, microsatellite analysis, SNP genotyping
			3 hrs	900 bp†	1–96	0.9–86.4 kb	
Illumina MiSeq® System	50 ng Nextera® kit	1 lane flow cell	4 hrs	1 × 36 bp§	3.4 million (single reads)	1 Gb	DNA sequencing, gene regulation analysis, quantitative and qualitative sequencing-based transcriptome analysis, SNP discovery and structural variation analysis, cytogenetic analysis, DNA-protein interaction analysis (ChIP-Seq), sequencing-based methylation analysis, small RNA discovery and analysis, de novo, metagenomics, metatranscriptomics
	0.1–1 µg TruSeq® kit		27 hrs	2 × 150 bp**			
Illumina HiSeq® System	50 ng Nextera kit	Single or dual 8-lane flow cell	1.5–11 days	2 × 100 bp‡	3 billion (single reads)	Up to 600 Gb	
	0.1–1 µg TruSeq kit						

* CE-based run time does not include 4-hour sequencing reaction on the thermocycler prior to loading. NGS sequencing and base detection occurs concurrently during the run.
† Base pairs with quality scores of 20 (Q20).
§ > 90% of base pairs have quality scores of 30 (Q30).
** > 75% of base paris have quality scores of 30 (Q30).
‡ > 80% of base pairs have quality scores of 30 (Q30).

Tunable Resolution

NGS provides a high degree of flexibility for the level of resolution required for a given experiment. A sequencing run can be tailored to produce more or less data, zoom in with high resolution on particular regions of the genome, or provide a more expansive view with lower resolution. To adjust the level of resolution, a researcher can tune the coverage generated for a particular type of experiment. The term coverage generally refers to the average number of sequencing reads that align to each base within the sample DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was covered by 30 sequencing reads.

The ability to easily tune the level of coverage and resolution offers a number of experimental design advantages. For instance, in cancer research, somatic mutations may only exist within a small proportion of cells in a given tissue sample. Using mixed-cell samples, the region of DNA harboring the mutation must be sequenced at very high levels of coverage, upwards of 1000×, to detect these low frequency mutations within the cell population. While this type of analysis is possible with CE technology, there is an additive cost incurred with each additional read, so experiments requiring high read depths can become prohibitively expensive, especially when scaling the process across a number of samples.

On the other side of the coverage spectrum, a researcher would likely choose a much lower coverage level for an application like genome-wide variant discovery. In this case, it makes more sense to sequence at lower resolution, but process larger sample numbers to achieve greater statistical power within a given population of interest.

Unlimited Dynamic Range

The digital nature of NGS supports an unlimited dynamic range, providing very high sensitivity for quantifying applications, such as gene expression analysis. With NGS, researchers can quantify RNA activity at much higher resolution than traditional microarray-based methods, important for capturing subtle gene expression changes associated with biological processes. Where microarrays measure continuous signal intensities, with a detection range limited by noise at the low end and signal saturation at the high end, NGS quantifies discrete, digital sequencing read counts. By increasing or decreasing the number of sequencing reads, researchers can tune the sensitivity of the experiment to accommodate different study objectives.

Universal Biology Tool

The powerful and flexible nature of NGS has permeated many areas of study, becoming firmly entrenched as an indispensable and universal tool for biological research. With the ability to analyze the genetic architecture of any biological entity, the scientific community has used this technology platform to develop a broad range of applications that have transformed study designs, surpassing boundaries, and unlocking information never before imaginable.

Diverse Applications

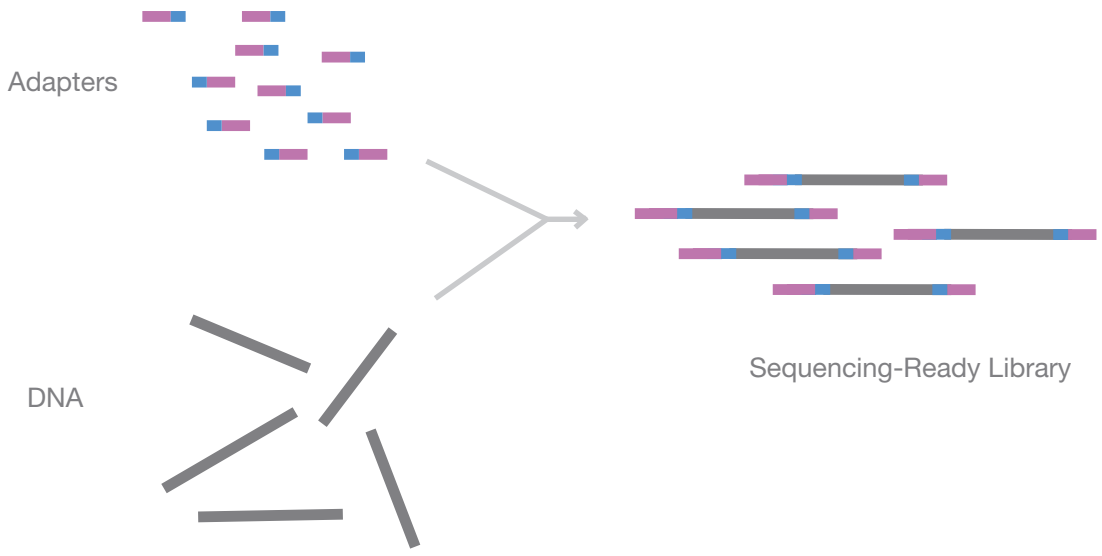
Next-generation sequencing (NGS) platforms enable a wide variety of applications, allowing researchers to ask virtually any question of the genome, transcriptome, and epigenome of any organism. Sequencing applications are largely dictated by the way sequencing libraries are prepared and the way the data is analyzed, with the actual sequencing stage remaining fundamentally unchanged. There are a number of standard library preparation kits that offer protocols for sequencing whole genomes, mRNA, targeted regions such as whole exomes, custom-selected regions, protein-binding regions, and more. To address specific research objectives, many researchers have developed novel protocols to isolate specific regions of the genome associated with a given biological function.

Sample preparation protocols for NGS are generally more rapid and straightforward than those for capillary electrophoresis (CE)-based Sanger sequencing (Table 2). With NGS, researchers can start directly from a gDNA or cDNA library. The DNA fragments are then ligated to platform-specific oligonucleotide adapters needed to perform the sequencing biochemistry, requiring as little as 90 minutes to complete (Figure 3).

Table 2: Sample Preparation for Whole-Genome Sequencing at a Glance

CE-based Sanger Sequencing	Next-Generation Sequencing
Library preparation more involved—each sample must contain a single template, requiring purification from single bacterial, yeast colonies, or phage plaques	Library preparation more streamlined—sample can consist of a population of DNA molecules that do not require clonal purification
Complete within days to weeks, depending upon the size of the genome being sequenced	Complete within hours, regardless of genome size

Figure 3: NGS Library Preparation



NGS library preparation starts directly from fragmented genomic DNA. Platform-specific oligonucleotide adapters, needed to perform the sequencing biochemistry and index samples, are ligated to each end of the fragments to yield the sequencing-ready library.

In contrast, CE-based Sanger sequencing requires genomic DNA to be fragmented first and cloned into either bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs). Then, each BAC/YAC must be further subcloned into a sequencing vector and transformed into the appropriate microbial host. Template DNA is then purified from individual colonies or plaques prior to sequencing. This process can take days or even weeks to complete, depending on the size of the genome.

Data Analysis Algorithms

Along with sample preparation, data analysis is an important factor to consider for sequencing applications. A range of data analysis algorithms are available that perform specific tasks related to a given application. Some applications, such as *de novo* sequencing, require specialized assembly of sequencing reads. Other applications, such as RNA-Seq, require algorithms that quantify read counts to provide information about gene expression levels. A number of algorithms exist to address the needs of each application. While some of these are commercially available from software vendors, many are freely available open-source algorithms from academic institutions. Illumina collaborates closely with commercial and academic software developers to create an ecosystem of data analysis tools that address the needs of various research objectives[1].

Whole-Genome Sequencing

Until recently, sequencing an entire genome was a major endeavor. Even for fairly compact viral genomes with overlapping genes and few to no repetitive regions, whole-genome sequencing using CE-based Sanger technology requires a significant commitment of time and resources. For example, *de novo* whole-genome sequencing of vaccinia virus—a large and complex DNA virus (~200 kb)—using CE-based methods would involve roughly 4,000 sequencing reactions (assuming 10× coverage and 500 bp read lengths), each conducted in separate tubes or wells. Using NGS technology, the same sequencing project, including library preparation, can be completed in just a few days with one sequencing run at 30× coverage or greater.

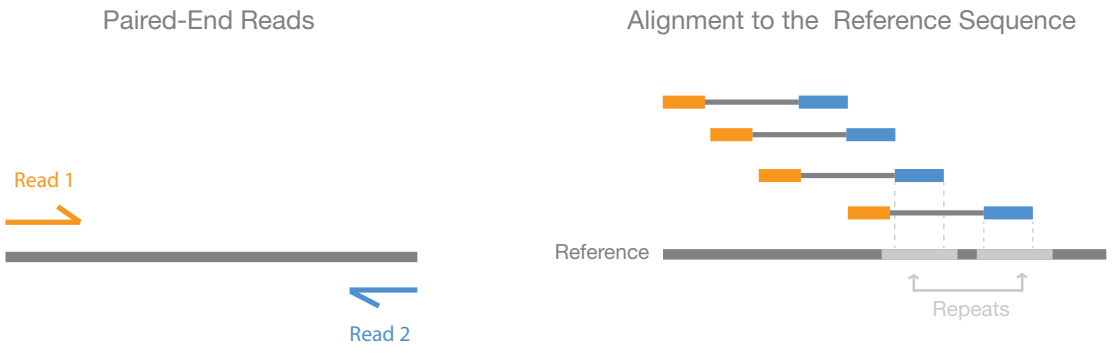
Sequencing Small Genomes

The ability of NGS platforms to produce a large volume of data in a short period of time makes it a powerful tool for whole-genome sequencing in a laboratory environment. While NGS is commonly associated with sequencing large genomes, especially human genomes, the scalability of the technology makes it just as useful for small viral or bacterial genomes. This was demonstrated during the recent *E. coli* (EHEC) outbreak in Europe, which prompted a rapid scientific response. Using the latest NGS systems, researchers were able to quickly generate a high-quality whole-genome sequence of the bacterial strain[2], enabling them to better understand the genetic mutations conferring the increased virulence.

One challenge associated with sequencing small genomes is the lack of reference genomes available for most species. This means that whole-genome sequencing must often be done *de novo*, where the reads are assembled without aligning to a reference sequence. The coverage quality of a *de novo* sequencing data set depends upon the quality of the contigs, or continuous sequences generated by aligning overlapping sequencing reads. The size and continuity of the contigs will affect the number of

gaps present in the data. A problem for *de novo* sequencing is that the short read lengths generated by NGS can lead to a higher number of gaps, regions where no reads align, resulting in greater fragmentation and smaller contigs—poorer data quality. This is especially true for regions of the genome containing repetitive sequence elements. To overcome this challenge, some NGS platforms offer paired-end (PE) sequencing protocols (Figure 4), where both ends of a DNA fragment are sequenced, as opposed to single-read sequencing where only one end is sequenced. Paired-end reads result in superior alignment across regions containing repetitive sequences and produce longer contigs for *de novo* sequencing by filling gaps in the consensus sequence, resulting in complete overall coverage.

Figure 4. Paired-End Sequencing and Alignment



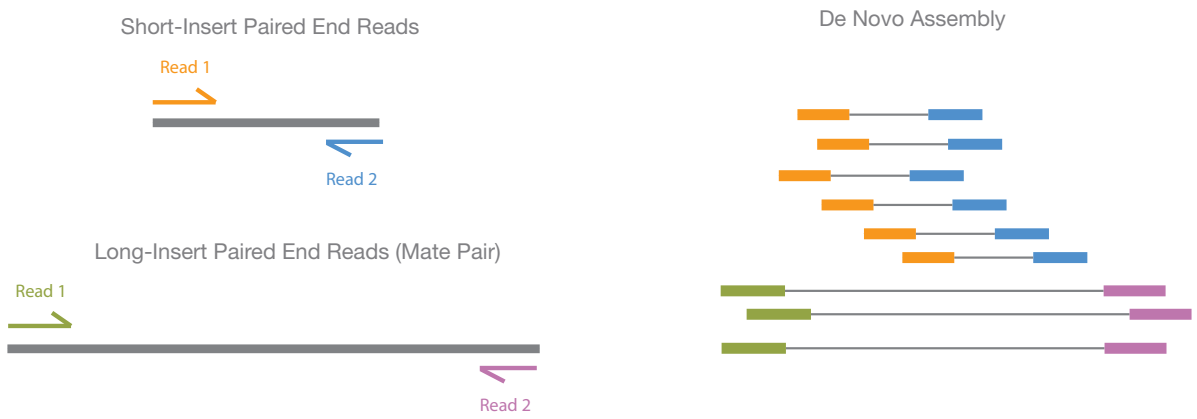
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Another important factor in generating high quality *de novo* sequences is the diversity of insert (DNA fragment) sizes in the library. Using longer inserts provides the highest fragment diversity relative to starting input material, yielding more uniform sequencing coverage. When long inserts are prepared for pair-end sequencing, a mate pair library is generated. These can include insert sizes ranging from 2 to 5 kb, optimal for *de novo* assembly applications, including both genome scaffold generation and genome finishing. In general, libraries with larger insert sizes will result in less fragmented assemblies and larger contigs. Combining short-insert paired-end and long-insert mate pair sequencing is the most powerful approach for maximal coverage across the genome (Figure 5). The combination of insert sizes enables detection of the widest range of structural variant types and is essential for accurately identifying more complex rearrangements, which results in a higher quality assembly. 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.

In parallel with NGS technological improvements, many algorithmic advances have been made in *de novo* sequence assemblers for short-read data. Researchers can perform high-quality *de novo* assembly using NGS reads and publicly available short-read assemblers. In many instances, existing computer resources in the laboratory are enough to perform *de novo* assemblies. The *E. coli* genome can be assembled in as little as 15 minutes using a 32-bit Windows desktop computer with 32 GB of RAM.

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTA
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Figure 5. *De Novo* Assembly with Mate Pairs



Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

Targeted Sequencing

With targeted sequencing, only a subset of genes or defined regions in a genome are sequenced, allowing researchers to focus time, expenses, and data storage on the regions of the genome in which they are most interested. This approach is typically used to sequence many individuals to discover, screen, or validate genetic variation within a population. The ability to pool samples and obtain high sequence coverage during a single run allows NGS to identify rarer variants that are missed, or too expensive to identify, using CE-based sequencing approaches. There are two different methods for making libraries for targeted sequencing and resequencing projects—target enrichment and amplicon generation (Table 3).

Table 3: Targeted Sequencing Sample Prep at a Glance

CE-based Sanger Sequencing	Next-Generation Sequencing
Library preparation more involved—each sample must contain a single template, either from a single PCR purified from single bacterial colonies	Library preparation more streamlined—each sample can be a population and does not require clonal purification
Suitable for sequencing amplicons and clone checking	Suitable for sequencing amplicons and clone checking
Complete within days to weeks, depending upon the size of the genome being sequenced	Complete within hours

Amplicon Sequencing

Amplicon sequencing allows researchers to sequence small, selected regions of the genome spanning hundreds of base pairs. The latest NGS amplicon library preparation kits allow researchers to perform rapid in-solution amplification of custom-targeted regions from genomic DNA. Using this approach, thousands of amplicons spanning multiple samples can be simultaneously prepared and indexed in a matter of hours. With the ability to process numerous amplicons and samples on a single run, NGS is much more cost-effective than CE-based Sanger sequencing technology, which does not scale with the number of regions and samples required in complex study designs. NGS enables researchers to simultaneously analyze all genomic content of interest in a single experiment, at fraction of the time and cost.

This highly targeted NGS approach enables a wide range of applications for discovering, validating, and screening genetic variants for various study objectives. For example, sequencing amplicons at a high depth of coverage can identify common and rare sequence variations. With sufficient coverage depth, deep sequencing can characterize rare sequence variants in a population, with minor allele frequencies below 1%. This application is particularly useful for the discovery of rare somatic mutations in complex samples (e.g., cancerous tumors mixed with germline DNA). Hence, amplicon sequencing is well-suited for clinical environments, where researchers are examining a limited number of disease-related variants.

Another common amplicon application is sequencing the bacterial 16S rRNA gene across a number of species, a widely used method for studying phylogeny and taxonomy, particularly in diverse metagenomic samples[3]. This method has been used to evaluate bacterial diversity in a number of environments, allowing researchers to characterize microbiomes from samples that are otherwise difficult or impossible to study. An overwhelming majority of the world's micorganisms have evaded cultivation, but sequencing-based metagenomic analyses are finally making it possible to investigate their ecological, medical, and industrial relevance[4].

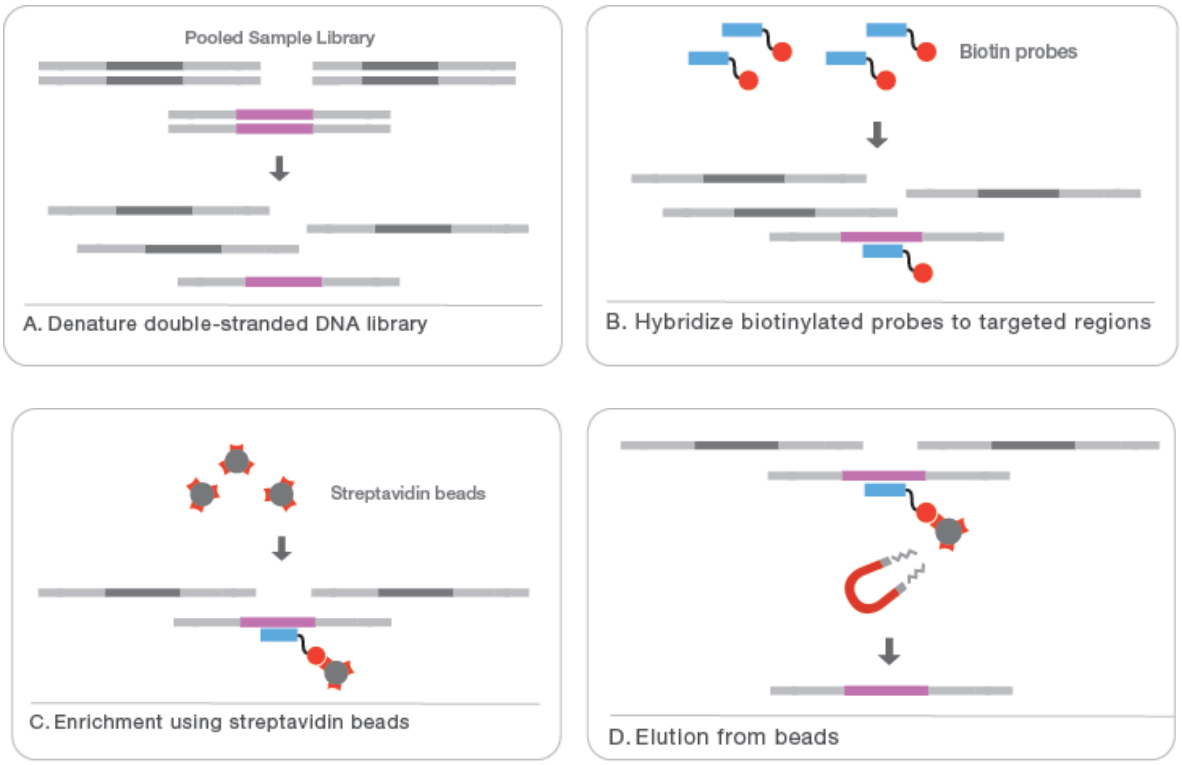
Target Enrichment

Target enrichment is similar to amplicon capture in that selected regions or genes are enriched in the library from genomic DNA. The difference is that target enrichment allows for larger DNA insert sizes and enables a greater amount of total DNA to be sequenced per sample. This ability lets researchers expand the information they can garner from each sample. For example, rather than sequencing a few hundred exons, they can sequence the entire exome to call functional single nucleotide polymorphisms (SNPs) for an individual. Sequencing human exomes from many individuals enables rare disease-associated alleles to be identified within a population.

Numerous kits exist for target enrichment. Researchers can choose to either design custom probes or use standard kits for widely used applications, such as exome enrichment. With custom design, researchers can target regions of the genome relevant to their research interests. This is ideal for examining specific genes in pathways, or as a follow-up analysis to genome-wide association studies (GWAS). Segments of the genome identified in a GWAS as harboring causative variants can be enriched and sequenced to identify additional, possibly rarer, variants in the associated region.

The steps involved in making a NGS library for target enrichment are virtually identical to the steps involved in library generation for whole-genome sequencing, with the addition of a target enrichment step (Figure 6). Genomic DNA is isolated, fragmented, purified to ensure uniform and appropriate fragment size, and end-ligated to oligonucleotide adapters specific to the NGS platform. The complexity of a sequencing library can then be reduced by enriching for specific regions of interest, such as exons or specific genes, by immobilizing DNA on a substrate such as magnetic beads. After binding the library and washing away unbound DNA, a simple elution step allows recovery of the library, which can then be amplified by PCR in preparation for sequencing.

Figure 6: Target Enrichment



In target enrichment, selected regions are captured from the library by probes bound to magnetic beads. Once the unbound DNA in the library is washed away, the capture DNA is eluted to provide an enriched library.

From Innovation to Publication

The advent of NGS has enabled researchers to study biological systems at a level never before possible. As the technology has evolved, an increasing number of innovative sample preparation methods and data analysis algorithms have enabled a broad range of scientific applications. Researchers are making fascinating discoveries in a number of biological fields, unlocking answers never before possible. As a result, there has been an explosion in the number of scientific publications. Illumina sequencing alone has resulted in thousands of peer-reviewed publications. Selected recent examples are listed below.

Whole-Genome Sequencing

- Srivatsan A, et al. (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet 4: e1000139.
- Rasmussen M, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757-762.
- Li R, et al. (2010) The sequence and *de novo* assembly of the giant panda genome. Nature 463:311-317.
- Pelak K, et al. (2010) The characterization of twenty sequenced human genomes. PLoS Genet 6:e1001111.

Targeted Resequencing

- Ram JL, Karim AS, Sendler ED, Kato (2011) Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the Illumina sequencing platform. Syst Biol Reprod Med. 57(3):117–8.
- McEllistrem M.C. (2009) Genetic diversity of the pneumococcal capsule: implications for molecular-based serotyping. Future Microbiol 4:857-865.
- Lo YMD, Chiu RWK. (2009) Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. Clin. Chem 55:607-608.
- Robinson PN (2010) Whole-exome sequencing for finding *de novo* mutations in sporadic mental retardation. Genome Biol 11:144.
- Araya CL, Fowler DM (2011) Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. doi:10.1016/j.tibtech.2011.04.003.

References

1. www.illumina.com/software/illumina_connect.ilmn
2. www.illumina.com/Documents/products/appnotes/appnote_miseq_ecoli.pdf
3. www.illumina.com/Documents/products/appnotes/appnote_miseq_denovo.pdf
4. Bomar L, Maltz M, Colston S, and Graf J. (2011) Directed culturing of microorganisms using metatranscriptomics. mBio. 2:e00012–11

FOR RESEARCH USE ONLY

© 2011-2012 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.

Pub No. 770-2012-008 Current as of 28 February 2012

illumina®