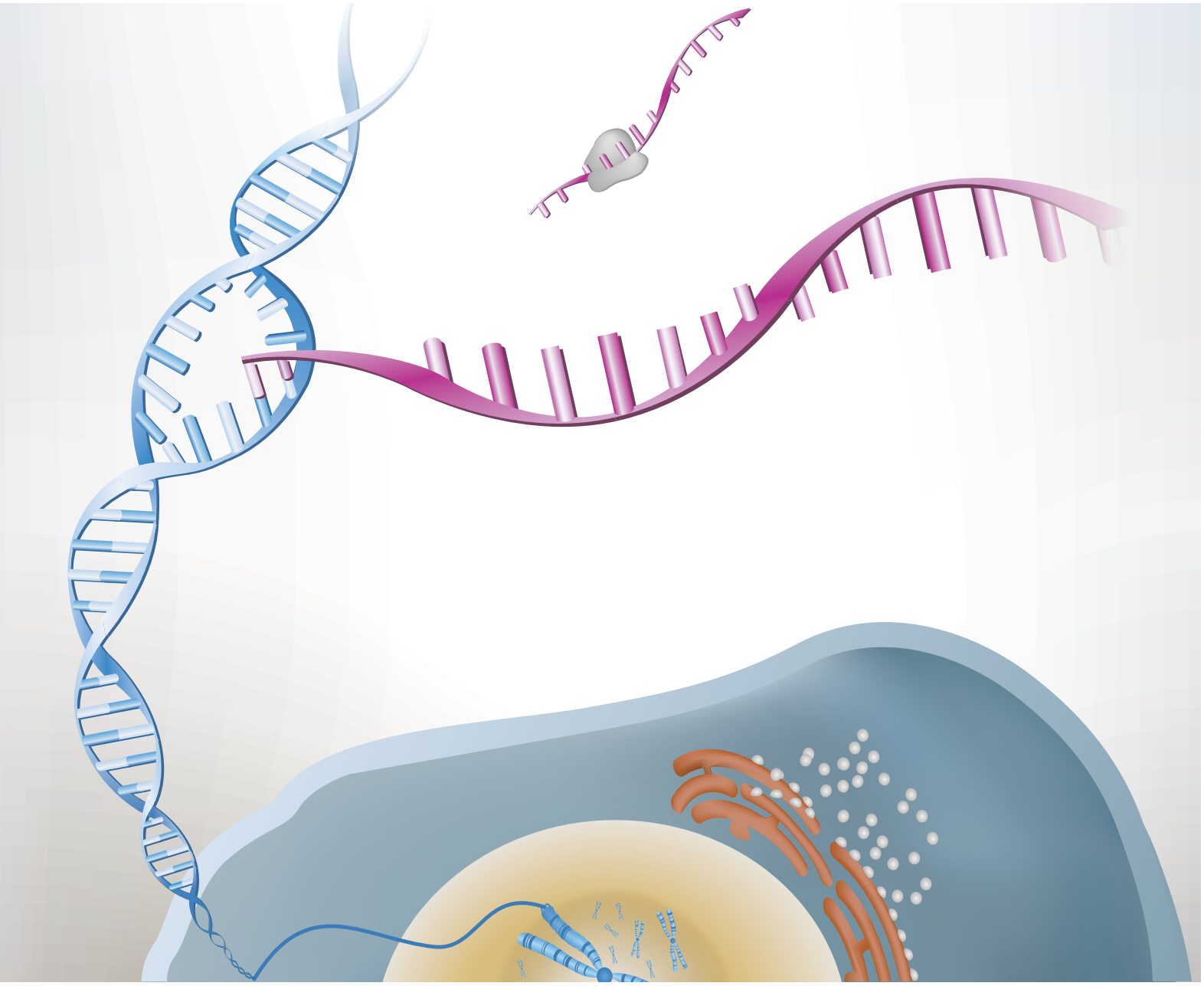
illumina®

An introduction to next-generation sequencing for cell biologists



www.illumina.com/NGStech

Table of contents

Welcome to next-generation sequencing 3

Why next-generation sequencing? 3

- High-throughput science 3
- Dynamic range 4

Sequencing applications for cell biology 4

- Gene expression profiling 5
- Total RNA and mRNA sequencing 5
- Targeted RNA sequencing 5
- Small RNA and noncoding RNA sequencing 5
- Characterizing DNA methylation 6
- Characterizing chromatin accessibility 6
- Understanding DNA-protein interactions 6

What is next-generation sequencing? 7

- The basic next-generation sequencing workflow 7
- Multiplexing 7
- Sequencing services and systems 8
- Single-cell analysis 8

Summary 9

Glossary 10

References 11

Welcome to next-generation sequencing

In 1665, Robert Hooke placed a thin slice of cork under his microscope and saw that it divided into "a great many little boxes," which he referred to as "cells." His findings and the invention of the microscope led to the cell theory, the foundation of biology. As the basic unit of life, cells play a role in all aspects of human physiology, including development and disease. At the time, the microscope revolutionized how scientists understand cellular function, and it has become an invaluable tool for studying the appearance and behavior of cells. However, new discoveries have revealed that cell activity involves more than physical characteristics alone.

Today, biologists studying disease etiology have many methods at their disposal to analyze cells—microscopy, flow cytometry, and staining comprise only a few. Many conventional approaches for studying cell biology rely on antibody staining or other signal-based methods to visualize protein interactions. Using staining alone to study protein function can be challenging, because antibodies are not available for all proteins and only a limited number of targets can be investigated simultaneously. These methods reveal physical attributes of cells and protein function, but do not capture an important component of biology—namely, the genetic code that dictates protein behavior. Increasingly, scientists are finding strong links between genetics and disease phenotypes, indicating that a true understanding of proteins and their behavior requires knowing the underlying genetics and regulation.

Understanding genetic factors is essential for virtually all branches of biological research, and the introduction of next-generation sequencing (NGS) technology has transformed the study of biology.[1] Recent collaborations such as the Encyclopedia of DNA Elements Consortium (ENCODE)[2] used NGS to provide functional information about human health at the molecular level by exploring transcription, regulation, DNA-protein binding, and epigenetics. By digitally counting sequence reads rather than measuring continuous signal intensities, as many current methods do, NGS provides quantitative information about protein translation, gene expression, and regulation at higher resolution than traditional microarray—or antibody-based—methods. Knowing how these processes are altered in disease states can help researchers predict changes to cell behavior and design additional studies to assess the proteins in specific functional pathways. NGS opens new avenues to exploring and understanding the cellular activity of disease, so scientists can ultimately develop therapeutic approaches to target the biological pathways contributing to disorders.

Why next-generation sequencing?

As an unbiased technology, NGS can be used to investigate gene expression and protein translation in disease biology and generate hypotheses for further functional studies. It offers a high-throughput alternative to current single-analyte assays, enabling screening of more samples in less time. Using NGS, biologists can analyze cell states and fates, study signal transduction pathways, determine causal variants, and examine tissue-specific gene expression.

High-throughput science

Often, biological research involves single-gene knockout models that assess how the absence or modification of a certain protein affects a specific pathway. This approach is time-consuming and does not always result in conclusive findings, but often leads to additional experiments designed to test alternate hypotheses. Today, many scientists are using other approaches to find targets of interest. NGS enables researchers to survey the entire genome, transcriptome, or epigenome in a single sequencing run and observe multiple changes at one time. Sequencing results can inform experimental design by identifying proteins of interest for subsequent studies, so that researchers can investigate targets efficiently, save time, and publish sooner.

Dynamic range

Dynamic range is the difference between the highest and lowest signals that can be detected. The digital nature of NGS supports large dynamic range, providing high sensitivity for quantification-based applications, such as gene expression analysis. With NGS, researchers can quantify RNA activity at much higher resolution than traditional microarray-based methods, which is important for capturing subtle gene expression changes associated with biological processes.[3] While 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 an experiment to accommodate different study objectives.

Sequencing applications for cell biology

NGS platforms enable a wide variety of applications, allowing researchers to investigate any cell type or organism. In disease cases, various factors can influence the affected phenotype, including gene mutations, changes to gene expression at different times and under varying environmental conditions, and differences in epigenetic regulation. Cell biologists must consider these potential alterations when looking for the molecular pathways that contribute to complex traits or diseases. Researchers can leverage the flexibility of NGS to quantify gene expression in a signaling pathway using a single technology (Figure 1).

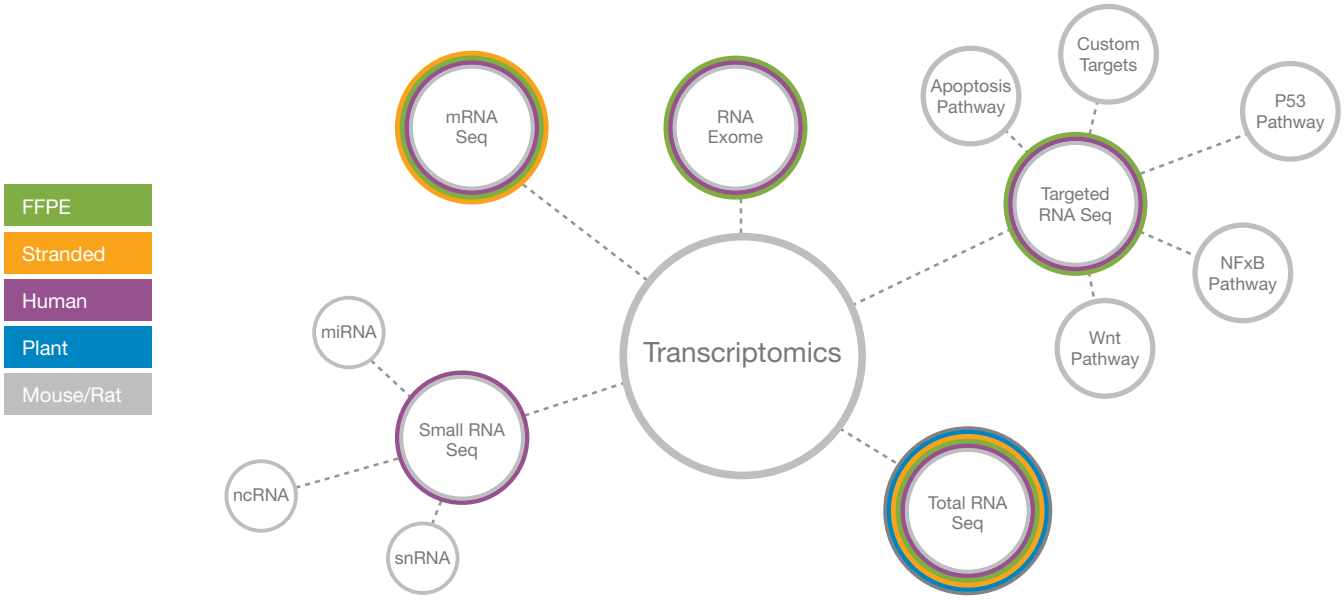


Figure 1: A Complete View of Transcriptomics with NGS—A broad range of methods for transcriptomics with NGS have emerged over the past 10 years including total RNA-Seq, mRNA-Seq, small RNA-Seq, and targeted RNA-Seq.

Gene expression & regulation profiling

Traditional approaches for evaluating gene expression and regulation include microarrays and quantitative polymerase chain reaction (qPCR) which require prior knowledge of probe design. RNA sequencing (RNA-Seq) is a more sensitive approach that can quantify RNA species by measuring discrete, digital sequencing reads.[4] Library preparation methods for RNA-Seq typically begin with total RNA sample preparation, which is then converted into cDNA prior to standard NGS library preparation. RNA-Seq focused on mRNA, small RNA, non-coding RNA, or microRNAs can be achieved by including additional isolation or enrichment steps before cDNA synthesis (Figure 1).

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, a large 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.[5] 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 to detect differential expression, allele specific expression, detection of gene-fusions, isoforms, cSNPs, and splice junctions.

Small RNA and noncoding RNA sequencing

Small, noncoding RNA, or microRNA s 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.[6,7]

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.

Epigenetics is the study of heritable changes in gene activity caused by mechanisms other then DNA sequence changes. Recent studies have shown that lifestyle and environmental factors can lead to epigenetic changes to DNA, which can cause or exacerbate disease. Mechanisms of epigenetic activity include DNA methylation, DNA-protein interactions, chromatin accessibility, histone modifications and more.

Characterizing DNA methylation

Aberrant DNA methylation and its impact on gene expression have been implicated in many disease processes, including Alzheimer's disease.[8] Methylation of DNA at cytosine nucleotides impacts various cellular activities involving gene expression, RNA processing, and protein function. These processes can influence cell biology by activating or suppressing certain genes at specific times, with implications in development, disease, aging, and immune defense. 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. Additionally, targeted Methyl-Seq offers a cost-effective choice between whole-genome bisulfite sequencing and methylation arrays. 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

Characterizing chromatin accessibility

Eukaryotic genomes are packaged into chromatin, and how chromatin is packaged plays a key role in gene regulation and, ultimately, cell phenotype.[9] Binding of transcription factors to regulatory sequences on DNA requires that chromatin be in an open conformation. Traditional methods of assessing open chromatin states such as the DNAase I hypersensitivity assay can be time consuming and requires a large amount of cells.
The assay for transposase accessible chromatin using sequencing (ATAC-Seq) is a fast and robust method that allows for profiling of single cells or bulk samples.[10,11] In this assay, genomic DNA is exposed to Tn5 (a highly active transposase) that preferentially inserts into open chromatin sites and adds sequencing primers. The sequenced DNA identifies the open chromatin and data analysis can provide insight into gene regulation. ATAC-Seq has been used to better understand gene regulation in embryonic development, T-Cell activation, and cancer.[12,13]

Understanding DNA-protein interactions

Chromatin immunoprecipitation sequencing, or ChIP-Seq, can be used to survey interactions between proteins, DNA, and RNA. ChIP-Seq using NGS enables researchers to identify the binding sites of multiple protein targets, including transcription factors and histones, across the entire genome. Analysis of DNA–protein interactions can provide insight into the regulation of events that are essential for many biological processes and disease states. Current methods for transcription factor analysis, such as arrays and qPCR, are limited in scope and provide information about only a subset of genes. NGS allows a broader scope of analysis, enabling identification and investigation of the many genes potentially activated by transcription factors in disease states. With ChIP-Seq, researchers can better understand how chromatin modifications and local structural changes impact transcription factor activity in a cell and within signaling pathways.

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

What is next-generation sequencing?

NGS is an advanced technology that uses fluorescence to provide a base-by-base view of the genome, transcriptome, or epigenome. Illumina sequencing chemistry, sequencing by synthesis (SBS), is the most widely adopted and published NGS technology worldwide.[14] SBS detects single bases as they are incorporated into growing DNA strands.

The basic next-generation sequencing workflow

In principle, NGS is similar to Sanger (capillary electrophoresis–based) sequencing. The bases of a DNA or RNA fragment are sequentially identified from signals emitted as each fragment is resynthesized from a template strand. NGS scales up this process; millions of reactions occur in a massively parallel fashion, rather than being limited to a single or a few fragments. This advance enables rapid sequencing of large stretches of genomic information.
To illustrate how this process works, consider a single DNA or RNA sample. The nucleic acid is first fragmented into a library of smaller segments that can be sequenced efficiently. The newly identified strings of bases, called reads, are then reassembled using a known reference genome as a scaffold (resequencing). The full set of aligned reads reveals the entire sequence of the sample (Figure 2).

A detailed animation of Illumina sequencing is available at www.illumina.com/SBSvideo.

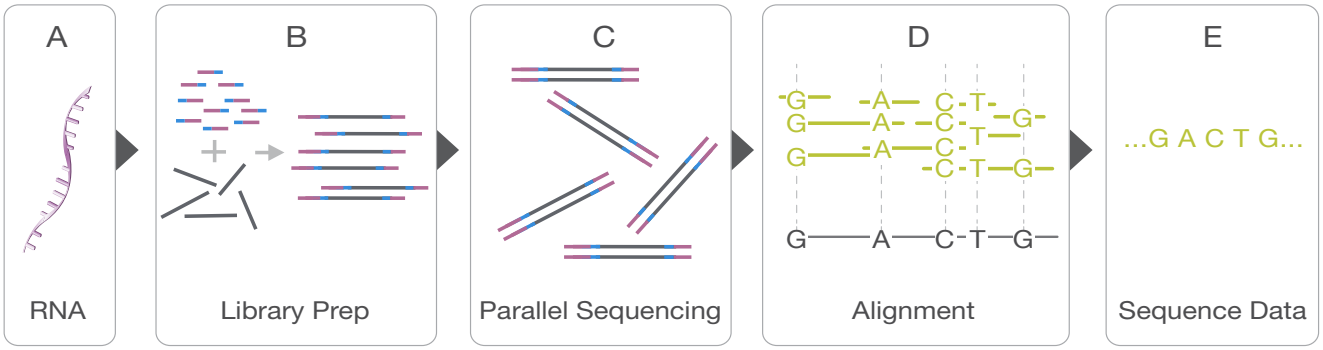


Figure 2: Concepts of NGS.

A. Nucleic acid (DNA or RNA) is extracted. If sequencing RNA, cDNA is reverse transcribed from the RNA template.
B. Library preparation fragments the DNA and adds adapters to the ends.
C. Fragments within the library are each sequenced in parallel.
D. Individual sequence reads are aligned to a reference sequence, such as a reference genome or known genes.
E. A consensus of aligned reads is generated allowing for subsequent variant calling.

Multiplexing

In addition to the increased amount of data generated on NGS instruments, the sample throughput per run has also increased over time. Multiplexing allows large numbers or batches of samples to be pooled and sequenced simultaneously during a single sequencing run (Figure 3). With multiplexed samples, unique index sequences (or "barcodes") are added to each DNA fragment during library preparation so that each read can be identified and sorted before final data analysis. Multiplexing with NGS allows investigators to process large sample numbers quickly, generating the statistical power to detect expression patterns confidently. Multiplexing dramatically reduces the time to data for multi-sample studies and enables researchers to go from experiment to answer faster and easier than ever before.

Sequencing services and systems

Service laboratories are a cost-effective option for researchers getting started with NGS and are located worldwide with sequencing services from many sample types. Providers offer services to accommodate diverse study designs, including RNA-Seq, bisulfite sequencing, ChIP-Seq, small RNA analysis, and others. Some service providers also deliver analyzed data, so that researchers spend less time analyzing data and can focus on the next discovery.

To locate a nearby service provider, contact your local account manager or contact Illumina customer service at customerservice@illumina.com.

Because of the advantages offered by NGS, many researchers choose to acquire their own NGS systems. Available systems range from high-output instruments that can process many samples in a single run to desktop sequencers ideal for smaller-scale studies.To find the sequencer that best fits your lab, visit www.illumina.com/sequencer.

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

Single-cell analysis

No one technology expanded the field of immunology more than flow cytometry. Flow cytometry gave researchers the ability to characterize immune cell phenotype at the single cell level, which led to our understanding of T-Cell and B Cell subsets and the specific role they play in immunity, tolerance, and tumorigenesis.[15] In contrast, genomic analysis techniques such as NGS, qPCR, and arrays are typically done on bulk samples of cells. Single-cell NGS is an emerging method that examines the genomes, transrciptomes, or epigenomes of individual cells, providing a high-resolution view of cell-to-cell variation in tissue. Single-cell sequencing can identify the cell in the context of its surrounding environment enabling researchers to assess cells individually rather than relying on the average signal from the entire population of cells. Today, there are over 75[16] single cell NGS methods developed on ILMN technology that allow researchers to profile the transcriptome[17], the epigenome[11], and genome[18] of single cells in a high throughput manner. The data from these studies is redefining our understanding of hematopoiesis[19], brain development[20] and tumorigenesis.[21]

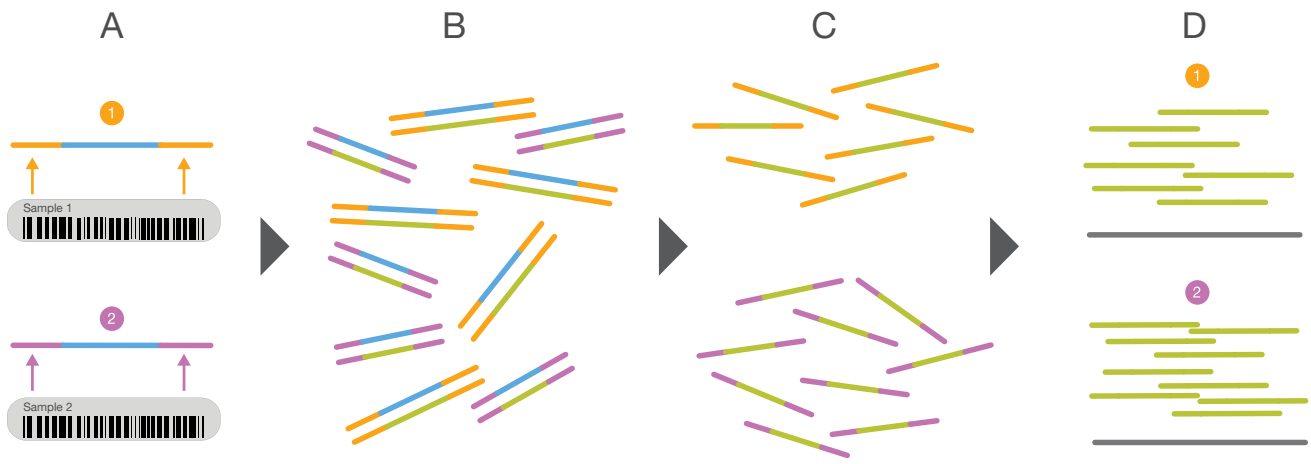


Figure 3: Sample Multiplexing Overview.

A. Two representative DNA fragments from two unique samples are 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.

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 understanding cellular function and variation. Investigating DNA and RNA sequences can provide insight into protein function and regulation, which has significant implications for disease research.

With NGS, scientists can perform multiplexed molecular analyses instead of sequential single-molecule analyses, establishing unbiased starting points for studies, progressing through research faster, and ultimately publishing sooner. The result is high-resolution, quantitative analyses to find signals from more protein precursors. As NGS is adopted into more laboratories and studies continue to map specific genetic profiles to phenotypes, more biologists than ever before have access to powerful genetic tools that can guide their experimental designs. Illumina is committed to providing the highest data quality in the industry, exemplified by implementation of the largest instrument install base of any NGS technology company[22] and relationships with leaders in many research fields. Together, we are bringing the promise of NGS toward a deeper understanding of human biology and disease.

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

Glossary

Adapters: Specialized oligos bound to the 5' and 3' ends 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 two distinct oligonucleotides. Repeated denaturation and extension cycles (similar to PCR) result in localized amplification of a single fragment into thousands of identical fragments. Millions to billions of unique, clonal clusters cover the flow cell. For Illumina NGS, cluster generation occurs on the sequencing instrument or in a separate fluidics instrument called a cBot.

Clusters: A clonal grouping of template DNA bound to the surface of a flow cell. Seeded by a single, template DNA strand, each cluster is clonally amplified through bridge amplification until the cluster has roughly 1,000 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 1-8 (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 384 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, and 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, contig assembly (*de novo* sequencing) must be used to construct the genome.

Sequencing by synthesis (SBS): SBS technology uses four 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.

References

1. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods*. 2008;5:16-18.
2. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011;9:e1001046.
3. Su Z, Labaj PP, Li S, et al. A comprehensive assessment of RNA-Seq accuracy, repodrucibility and information content byt he Sequencing Quality Control Consortium. *Nature Biotech*. 2014 32:903-914.
4. 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:e77910.
5. Wang Z, Gerstein M, Snyder M. RNA-Seq and revolutionary tool for transcritptomics. *Nature Rev Genet.* 2009; 10:57-63.
6. Wang Y, Kim S, Kim IM. Regulation of metastasis by microRNAs in ovarian cancer. *Front Oncol*. 2014;10:143.
7. 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.
8. Lunnon K, Smith R, Hannon E, et al. Methylomic profiling implicates cortical deregulation of *ANK1* in Alzheimer's disease. *Nat Neurosci*. 2014;17:1164-1170.
9. Kornberg, RD, Lorch Y. Chromatin structure and transcription. *Annu Rev Cell Biol*. 1992; 8:563-87.
10. Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNDNA-binding proteins and nucleosome position. *Nat Methods*. 10:1213-20.
11. Cusanovich DA, Daza R, Ade A, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. 2015; 348:910-14.
12. Cusanovich DA, Reddington JP, Garfield DA, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*. 2018; 555:538-542.
13. Gate RE, Cheng CS, Aiden AP, et al. Genetic determinants of co-accessible chromatin regions in activated T-cells across humans. *Nat Genet*. 2018; 50:1140-1150.
14. Data calculations on file. Illumina, Inc. 2017.
15. Practical Flow Cytometry. Fourth Edition. Shapiro, HM. 2013.
16. Uber Research, Dimensions analysis
17. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161:202-14.
18. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Research. 2018 25:1499-1507
19. Paul F, Ya'ara A, Giladi A, et.al.. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell 2015 163:1-15
20. Pollen AA, Nowakowski, TJ, Shuga J, et. al.. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotech. 2014 32:1053-61
21. Puram SV, Tirosh I, Parikh AS, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*. 2018;172:1-14.
22. Data calculations on file. Illumina, Inc. 2017.

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com
© 2018 Illumina, Inc. All rights reserved. PUB 1070-2014-006 QB 6452 Current as of 25 September 2018.

illumina®