illumina®

Multiomics: An overview of useful methods and applications

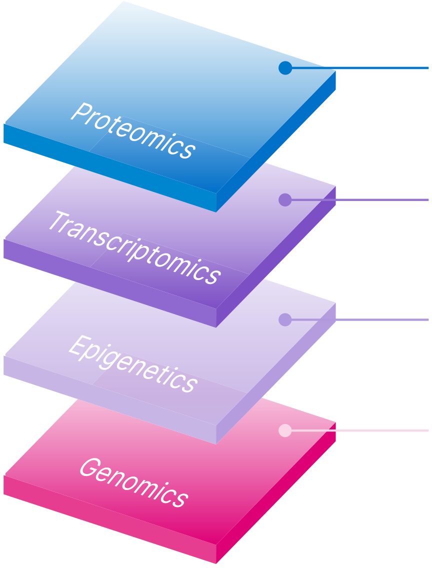
Harness the power of studying
multiple “omes” in one experiment

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

Table of contents

What is multiomics?

For over two decades, next-generation sequencing (NGS) technologies have revolutionized the understanding of life's complexity by enabling the comprehensive study of the genome, epigenome, transcriptome, and proteome. Such studies have also impacted the understanding of complex diseases and human health. Paradigm-shifting projects such as the Cancer Genome Atlas would not have been possible using legacy techniques which provide a limited view into complex questions. It is becoming clear that cells and systems are influenced by the complex interplay between several levels of biological regulation, and no single "ome" can fully explain biological phenomena. Thorough study of several levels of the central dogma (the movement of information from gene to protein) is therefore necessary to fully understand biological regulatory mechanisms and complex diseases.



Proteomics explores protein expression to understand cellular phenotypes and function.

Transcriptomics studies the complete set of RNA transcripts produced by a genome.

Epigenetics looks at how cells control gene activity through processes like DNA methylation.

Genomics focuses on the structure, function, evolution, mapping and editing of information in an organism's DNA.

Figure 1: Multiomics can be used to study biological omes in combination.

Multiomics uses NGS techniques to collect unbiased data from different biological levels—omes—in one experiment. Doing this allows researchers to examine the same problem from multiple angles.

Using multiomics, researchers can now explore complex questions by studying multiple modalities (data types) that include the genome, epigenome, transcriptome, and proteome—simultaneously by leveraging three broad experimental approaches: bulk-cell analysis, single-cell analysis, and spatial analysis.

A multiomics approach that leverages bulk-cell analysis will study the average gene or protein expression in pooled cell populations isolated from tissues. Bulk-cell analysis is the most straightforward and cost-effective multiomics approach. The bulk-cell approach also has a well-defined workflow that researchers who are new to NGS and multiomics can easily use. Furthermore, researchers can combine **bulk-cell analyses** with legacy experimental approaches like qPCR and flow cytometry to characterize cell populations of interest. For instance, researchers could use flow cytometry to sort different cell populations based on cell surface markers. Following the separation of these distinct cell populations, nucleic acid could be extracted from them which could be followed by bulk-cell sequencing. Specific gene targets uncovered by sequencing could then be further confirmed via qPCR.

While bulk-cell analysis has uncovered numerous valuable insights into enigmatic biological processes, it does not fully account for the inherent heterogeneity within cell populations. The approach also ignores the contributions of rare, yet biologically critical cell populations.

Single-cell analysis explores tissue heterogeneity by studying the genome, epigenome, transcriptome, or proteome of a single cell making it especially useful for studying rare cell populations. Cell "omics" can differ substantially within a microenvironment, even between cells that express the same biomarkers. Thus, studying the inherent individuality of cells and diversity of tissues can deliver useful insights researchers may miss with bulk-cell analysis.

While bulk-cell and single-cell analyses advance our understanding of biological processes at the molecular level, the data is gleaned outside of the context of the native tissue microenvironment of the cells. Experimental results from the two previous methods lose the context of the cell-cell interactions that happen in that microenvironment. The microenvironment and cell-cell communication have significant impact on biological function through gene expression and regulation. While it is possible to minimize the stress response of cells due to tissue dissociation and cell isolation, researchers have demonstrated that even "gentler" cell isolation processes can significantly affect the behavior of cells.

Furthermore, tissue imaging studies that have the goal of studying RNA and protein expression can be limited by various factors. Immunohistochemical experiments, for example, may allow researchers to look at only four or five proteins at once. Optimizing tissue samples for immunohistochemical experiments can be laborious and time-consuming and when they are successful, researchers are unable to gain a high throughput profile of the proteins that are expressed. Methods like fluorescence *in-situ* hybridization (FISH) are also limited in the breadth of parameters that can be studied in one experiment due to limitations on the number of dyes that can be used or by the quantity that the detection instrument can identify.

Spatial analysis allows researchers to study the genome and transcriptome within an unperturbed tissue microenvironment and with cell architecture maintained – a scenario that mimics realistic processes that happen in intact tissue systems.

Each of these approaches form the basis for the different multiomics methods we will discuss in this guide.

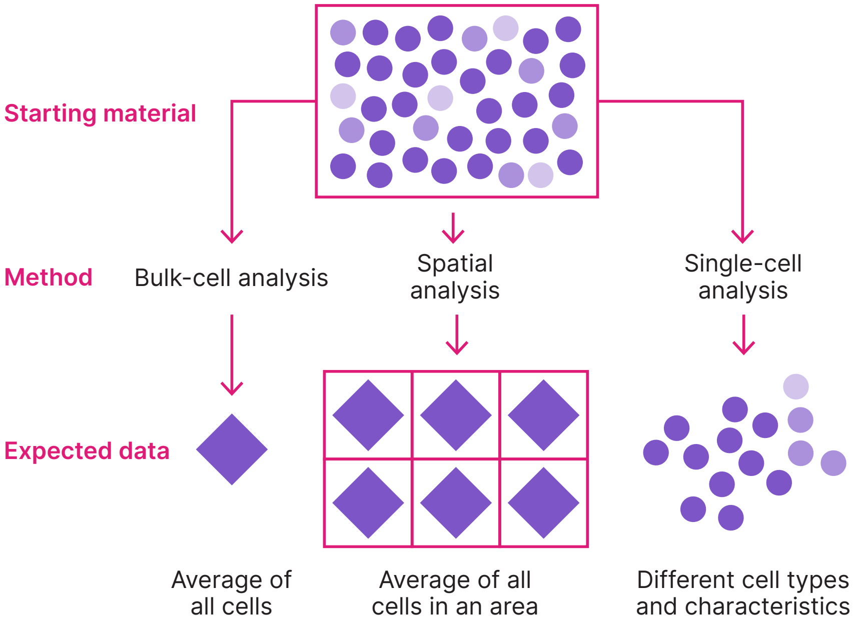


Figure 2: Multiomics leverages three broad experimental approaches: bulk-cell analysis, single-cell analysis, and spatial analysis.

Why multiomics?

Biological regulation is complex, and phenotypes are influenced by a myriad of factors that extend beyond DNA. An unbiased characterization of the cellular and environmental factors that influence genotypes and phenotypes is necessary to comprehensively understand novel systems.

Groundbreaking studies of individual omes have proven useful over the last century, providing life-saving treatments for complex diseases that were once considered untreatable. Notwithstanding, scientists still do not fully understand many biological processes. Multiomics allows researchers to combine the insights from multiple omes to fully understand fundamental biology and complex phenotypes.

Multiomics can be used in a myriad of biomedical disciplines, as we will discuss in the paragraphs below.

Multiomics in cancer research

Cancer is a disease of the genome, yet only 5%-10% of all cancers are a result of an inherited gene mutation.[1] This means approximately 90% of cancers cannot be fully explained by genetic mutations alone. Of the several mutations present in an average cancer, only a few of those mutations are “driver mutations” which promote tumorigenesis and give tumors a selective growth advantage.[2] The majority of mutations found in cancer are “passenger mutations” which don’t have identifiable mechanistic links to cancer initiation and progression. Yet while passenger mutations may not have characterized mechanistic links to cancer, they are not inconsequential.[3] Thus, though DNA sequencing is crucial to understanding cancer, profiling additional omes can provide significant insights into how cancer-associated genes are regulated and expressed.

Multiomics has also been a critical tool to provide researchers with unprecedented potential to understand why different patients with the same type of cancer have variable responses to the same treatment.[4] Single omics may not adequately explain these different response rates given that cancers contain heterogeneous populations of cells which change and evolve in response to treatment.[5] Given that those changes often manifest in multiple omes, multifactorial modeling using the genome, transcriptome, epigenome, and proteome is useful in accurately predicting and explaining divergent treatment response rates.

Multiomics enables the development of next-generation diagnostic tools

Diagnosing many childhood-onset neurological diseases is often challenging due to non-specific presentation or just a general rarity of the disease.[6] Multiomics is proving to be a paradigm-shifting approach that enables research and development of diagnostic tools that could unravel the medical enigma surrounding childhood-onset neurological conditions.[7,8,9] Up to 80% of rare diseases are genetic or have a genetic etiology.[10] Thus, while early diagnosis and intervention may not eliminate the disease, they could still help researchers find solutions that mitigate severe symptoms or presentations of the disease. Multiomics can uncover the DNA mutations as well as factors that influence gene expression underlying these diseases early in life, and thus prompt early intervention for the individual.

As the cost of omics technologies continues to decrease, more labs will be able to study diseases using multiomics sequencing. This may help physicians make quicker diagnoses and make timely decisions about a patient’s care.

75%

Cancer drugs can be ineffective in up to 75% of patients, highlighting a need for more thorough understanding of treatment response[11]

90%

of cancers cannot be fully explained by mutations alone

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

Multiomics in population-level studies

Sequencing biological samples from individuals with a specific disease or who belong to a particular ethnic group improves how researchers understand the genetic variants that influence common diseases. Doing these types of studies have led to the development of "biobanks" around the world.

Biobanks are primarily university-based repositories that collect and store biological samples (and sometimes data) that researchers can study to draw links between genotypes and phenotypes common to a population.

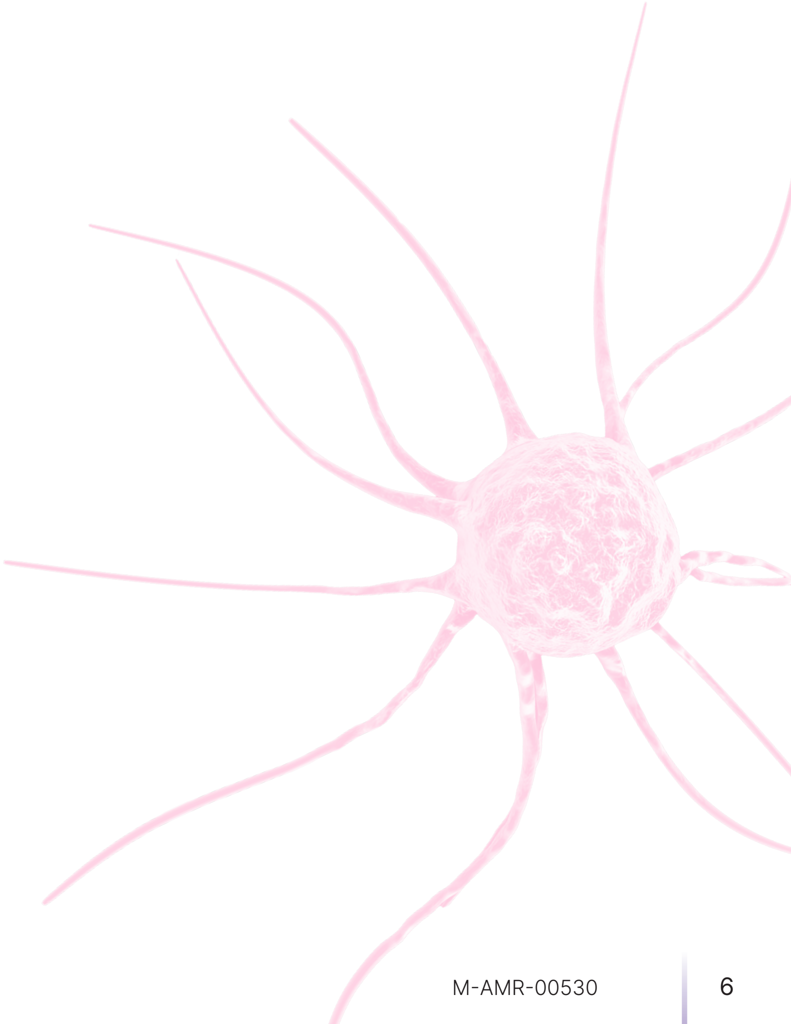
These insights can help researchers to:

- Find novel drug targets
- Identify patients most likely to benefit from a particular treatment approach
- Develop personalized treatment regimen for specific ethnic groups and
- Model longitudinal disease progression

It is worth noting, however, that there are drawbacks to biobanks. Historically, volunteers tend to be healthier and wealthier than the national average and don't represent the diversity of the population.[12,13] While several reasons may account for this, it doesn't minimize the power of biobanks and as such, researchers should take this disparity into consideration as they embark on multiomics-based population-level studies. An excellent example of a large biobank that is longitudinal and contains more diverse biological samples is the National Cancer Institute's Cancer Moonshot.

Multiomics in infectious disease research

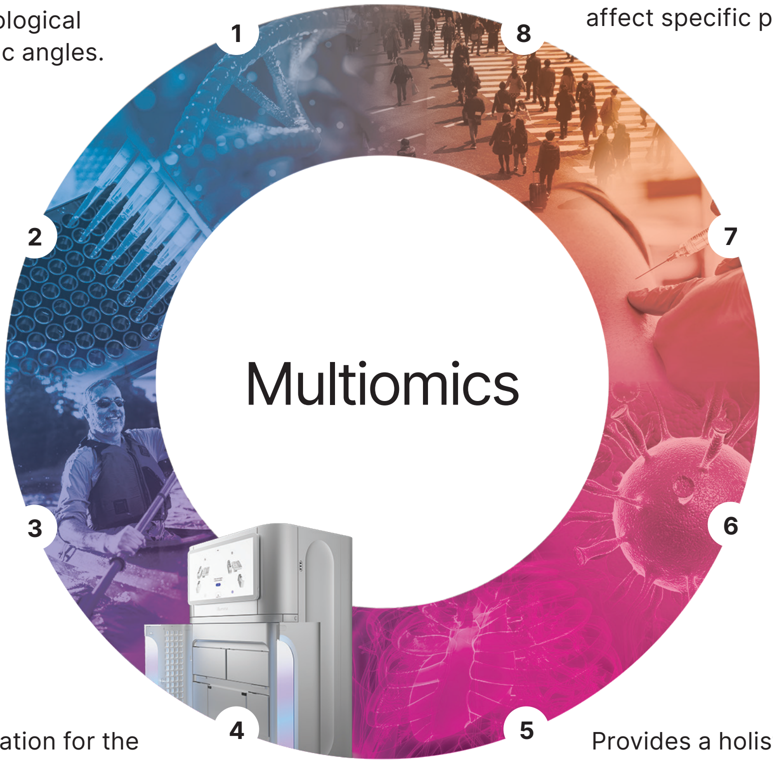
Researchers have used proteomic analyses to study the disease pathways of impactful human viruses like the Ebola virus[14], Zika virus[15] and the influenza virus[16]. Most recently, NGS analysis of SARS-CoV-2 has allowed researchers to understand the pathophysiological mechanisms of the virus in humans.[17] First, NGS helped scientists to rapidly identify the causative agent behind COVID-19.[18] NGS was also an instrumental tool that helped scientists understand the origins of the virus.[19] Furthermore, the availability of the SARS-CoV-2 sequence helped with the rapid development of diagnostic tests[20] and was a powerful trigger for the research on vaccines and therapeutics for COVID-19. NGS continues to help scientists track mutations in the virus.[21] On a much broader scale, looking at how SARS-CoV-2 affects multiple omes provides a framework that will not only be useful for this pandemic, but also in combating future epidemics.



Multiomics in neuroscience

Although researchers have studied the tau protein, β-amyloid and inflammation basis for Alzheimer’s Disease (AD) extensively over the years, the molecular drivers of this complex neurological disease are little understood. Currently, medications approved for AD only treat symptoms of the illness but do not cure it.[22] To gain a deeper understanding of the molecular changes that happen in AD, researchers used single-cell RNA sequencing and single-cell assay for transposase-accessible chromatin sequencing (ATAC-Seq)—a method that profiles open chromatin regions—to study 191,890 nuclei in late-stage AD and cognitively healthy controls from human brains postmortem. Using the ATAC-Seq datasets, researchers were able to identify cell-type-specific candidate cis-regulatory elements (cCREs) based on chromatin accessibility. They also found disease-associated cell subpopulation-specific transcriptomic changes, as well as transcription factors that may be regulating AD gene expression changes.[23]

In summary, multiomics is essential because it:



1. Is an unbiased tool that helps researchers understand the complex biological processes from various omic angles.
2. Explores the genomic, transcriptomic, epigenomic, and proteomic factors that underlie common, rare, and difficult-to-understand diseases.
3. Unravels therapeutic targets that have the potential to improve the lives of patients.
4. Forms a powerful foundation for the development of diagnostic tools that may detect diseases earlier than current detection methods and therefore prompt early intervention.
5. Provides a holistic biological view of mechanisms that govern health and disease.
6. Can help scientists respond quickly to epidemics.
7. Deepens our understanding of how these populations respond to medication.
8. Provides insights into how diseases affect specific populations.

Figure 3: Multiomics is an essential tool for studying various biological questions.

Frequently used multiomics combinations

Depending on your research goals, you will be able to choose a multiomics combination that suits your needs. In the paragraphs below, we provide information on frequently used multiomics combinations and the potential implications for your research. These are summarized in Table 1 on page 11.

Genomics and transcriptomics

DNA analysis allows for the detection of heritable markers that might offer information on an organism's susceptibility to a disease while incorporating RNA sequencing reveals the functional consequences of the genotype. By looking at both RNA and DNA, researchers maximize the likelihood of revealing therapeutically useful biomarkers.

Predicting coronary artery disease risk with genomics and transcriptomics

> Approximately 18 million people[24] die worldwide from cardiovascular diseases each year, yet scientists still know little about the molecular causes for these conditions.

In a recent study, researchers used NGS to profile the whole-genome and transcriptome of human coronary artery smooth muscle cells (HCASMCS), a cell type associated with coronary artery disease (CAD), from 52 unrelated donors. Genome data identified five genes *SIPA1, TCF21, SMAD3, FES* and *PDGRFRA*—all genes that play functional roles in vascular remodeling—that modulate CAD risk through HCASMCS. Additionally, mRNA sequencing further revealed that genes that were highly expressed in HCASMCS are enriched for single nucleotide polymorphisms (SNPs) known to be associated with CAD risk. These insights, along with epigenomic data that was collected during the study, allowed the researchers to create a map of gene regulation and expression in HCASMCS and how these contribute to CAD.[25]

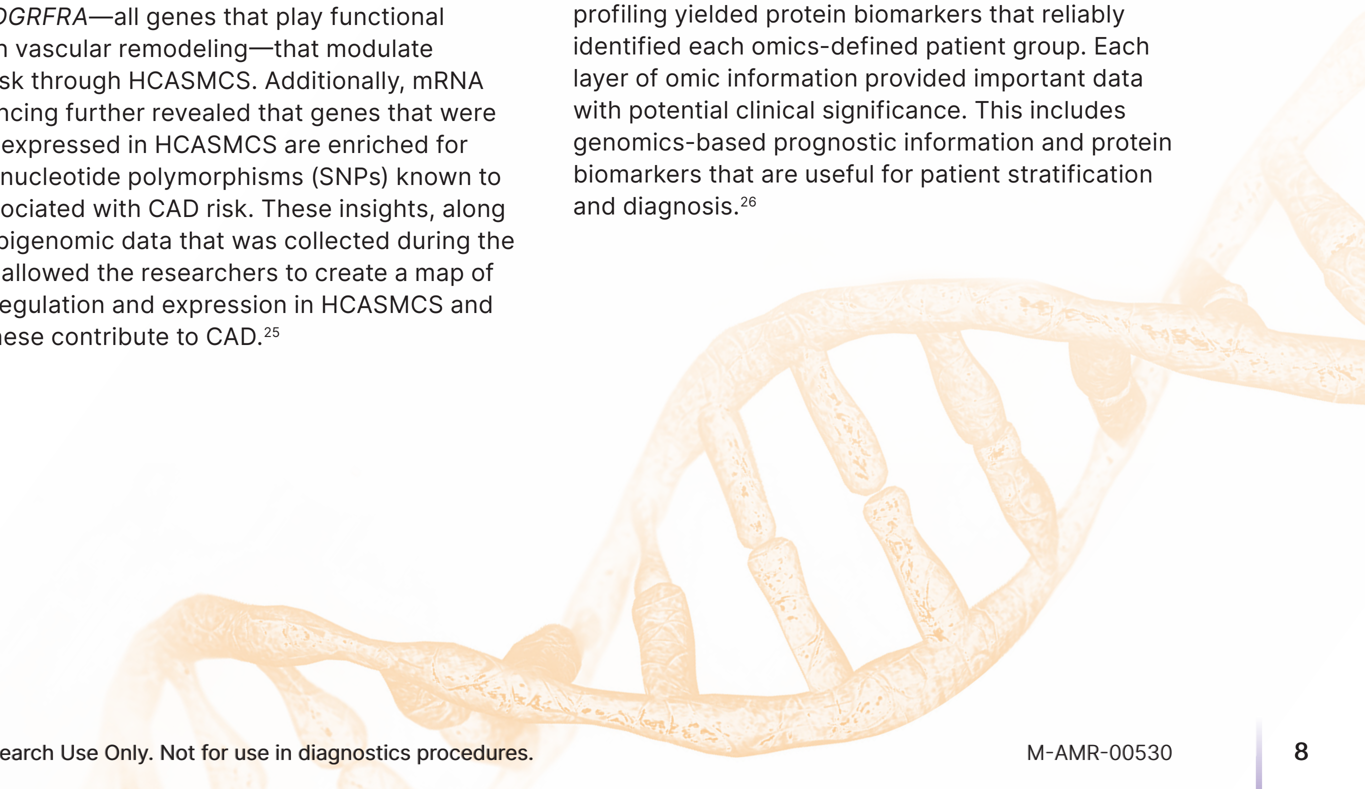
Genomics and proteomics

Several transcriptional and post-translational modifications occur as the genetic information stored in DNA is translated into protein. Since proteins often serve as biomarkers for illnesses like cancer, combining genomics with proteomics connects the genotype to a phenotype for in-depth learning into a disease state.

Proteogenomics in meningiomas

Meningiomas are the most common type of brain tumor, yet they are primarily treated surgically due to few effective medicinal therapeutics and difficulty in stratifying tumor types. To address historic difficulty in classifying and treating patients, researchers performed multiomic analyses on over 200 patient tumor samples. Genomic profiling of tumors, along with epigenomic and transcriptomic analyses, identified four patient groups with distinct prognoses.

Using RNA sequencing, the authors also identified upregulation of histone deacetylase activity (HDAC) in one of the patient groups and used animal models to demonstrate viability of HDAC inhibition as a possible treatment option. Finally, proteomic profiling yielded protein biomarkers that reliably identified each omics-defined patient group. Each layer of omic information provided important data with potential clinical significance. This includes genomics-based prognostic information and protein biomarkers that are useful for patient stratification and diagnosis.[26]



Transcriptomics and proteomics

RNA sequencing interrogates gene expression patterns that differentiate cells or distinct cell populations. Incorporating protein detection can link cell-specific expression with protein biomarkers. A key advantage of combining transcriptomics with proteomics is that it allows the combination of hypothesis-based panels of antibodies that interrogate proteins of interest with hypothesis-free analysis of the transcriptome for exploration and discovery in a single dataset.

Bulk epitope and nucleic acid sequencing (BEN-Seq) is a bulk-cell sequencing approach that uses oligo-conjugated antibodies that have been incorporated into an RNA sequencing workflow for simultaneously detecting RNA and protein data using NGS technology. Cellular indexing of transcriptomes and epitopes by sequencing (CITE-Seq) is another method that interrogates the protein and RNA profile of cells, but at the resolution of a single cell.

Studying host-pathogen interactions in COVID-19 with transcriptomics and proteomics

In response to an infection, interferon-stimulated genes (ISGs) are activated to produce chemical mediators that are highly effective at resisting and controlling pathogens. Researchers used CITE-Seq to study peripheral blood mononuclear cells (PBMCs) from 7 COVID-19 patients and 5 healthy donors. In the samples from the COVID-19 patients, the researchers noted an enriched expression for ISGs in T-cells and monocytes. This enhanced genetic signature appeared during early stages of COVID-19 infection and was not present in healthy donors.

In addition to the enhanced ISG signature, CITE-Seq revealed that there was a reduction of specific immune cell marker proteins in the presence of severe COVID-19 disease.[27] Understanding how genes and proteins are modulated differently in the presence of an infection provides unique insights into how the host immune system responds to pathogens and could lead to the discovery of druggable targets.

Genomics and epigenomics

Combining epigenomic findings with genomic information allows researchers to understand patterns of gene regulation and how they connect with the genotypes related to biological processes as well as diseases.

Predicting melanoma prognosis and overall survival with genomics and epigenomics

> Melanoma cases are detected in over 100,000 patients in the United States each year, many of which are treated with checkpoint inhibitors and immunotherapies.[28]

While these inhibitors generate a robust anti-tumor response in some patients, the mechanisms by which some overcome these novel therapeutics is poorly understood.

Multiomic modeling of therapeutic responses in a recent study found that no single gene or gene signature could explain melanoma treatment response rates.[3] However, a multidimensional model using data from whole-genome sequencing, RNA sequencing, and Illumina Infinium™ MethylationEPIC array (to profile methylation patterns) was able to predict treatment response with high sensitivity. The researchers found that three factors from the three omes were necessary to predict good response: high mutational burden, expression of the IFNγ response pathway (indicating an immune response), and low methylation of *PSMB8* (a gene encoding a proteosomal subunit). Conversely, poor response was associated with genomic structural variation, no enrichment of the IFNγ response pathway, and methylation of *PSMB8*. Alone, each omic lacked the power to predict why response rates were so different, but combining the analysis of the genome, transcriptome, and epigenome in one model increased predictive power and yielded insights that were previously unattainable.

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

Transcriptomics and epigenomics

Researchers studying the transcriptome and epigenome together allows them to directly measure the ties between gene regulation and gene expression.

Understanding the regulation of hematopoiesis during human development

To understand the regulation of hematopoiesis during human development, researchers used single-cell RNA sequencing (scRNA-Seq) and ATAC-Seq to study over 8000 matched human hematopoietic stem and progenitor cells (HSPCs) from fetal liver and bone marrow. Doing this allowed the researchers to identify 10 transcriptionally defined cell populations which had not been previously described in these fetal HSPCs. Using multiomics, researchers found that within transcriptionally homogeneous stem and progenitor cells, there are multiple subpopulations that differ in their overall chromatin assembly and lineage-specific transcription factor activity. This analysis led the researchers to conclude that in hematopoietic stem cells, programs that control the different fates of the cell are primed at the chromatin level, prior to their commitment to a specific lineage.[29]

Table 1: Frequently used multiomics combinations, implications for researchers and example applications.

Multiomics combination	Implications	Reference for application example
Genomics + Transcriptomics	An organism's genotype offers information on the susceptibility to a specific disease. Incorporating RNA sequencing helps researchers measure the functional consequences of genetic variation and may reveal biomarkers that are therapeutically useful.	Liu B, et al. Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. *Am J Hum Genet*. 2018;103(3):377-388.
Genomics + Proteomics	Combining genomics with proteomics connects the genotype to a phenotype for more in-depth understanding into a disease state.	Demaree B, et al. Joint profiling of DNA and proteins in single cells to dissect genotype-phenotype associations in leukemia. *Nat Commun*. 2021;12(1):1583.
Genomics + Epigenomics	Combining genomic findings with epigenomic information allows researchers to understand patterns of gene regulation and how they connect with the genotypes that underlie diseases.	Newell F, et al. Multiomic profiling of checkpoint inhibitor-treated melanoma: Identifying predictors of response and resistance, and markers of biological discordance. *Cancer Cell*. 2022;40(1):88-102.e7.
Transcriptomics + Proteomics	RNA sequencing interrogates gene expression patterns that differentiate cells or distinct cell populations. Incorporating protein detection can link cell-specific expression with protein biomarkers.	Arunachalam PS, et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science*. 2020;369(6508):1210-1220.
Transcriptomics + Epigenomics	Studying the transcriptome and epigenome together allows researchers to directly measure the ties between gene regulation and gene expression.	Ranzoni AM, et al. Integrative single-cell RNA-Seq and ATAC-Seq analysis of human developmental hematopoiesis. *Cell Stem Cell*. 2021;28(3):472-487.e7.

What do multiomics workflows look like?

Every multiomics experiment starts with the extraction of nucleic acids from cell, tissue, or liquid biopsy samples. Once nucleic acids have been isolated, libraries are prepared. A library is a collection of similarly sized nucleic acid fragments that have adapter DNA sequences added to the 5′ and 3′ ends. These adaptors are necessary because they make libraries compatible with Illumina sequencers. Following library preparation, researchers sequence the libraries, collect, and analyze the data. The method a researcher uses to isolate nucleic acids and/or prepare libraries will depend on the multiomics method they use. Regardless of the method, however, multiomics workflows follow the general sequence shown in Figure 4.

In this chapter, we provide a brief overview of sample prep/library preparation, sequencing, and multiomics data analysis.

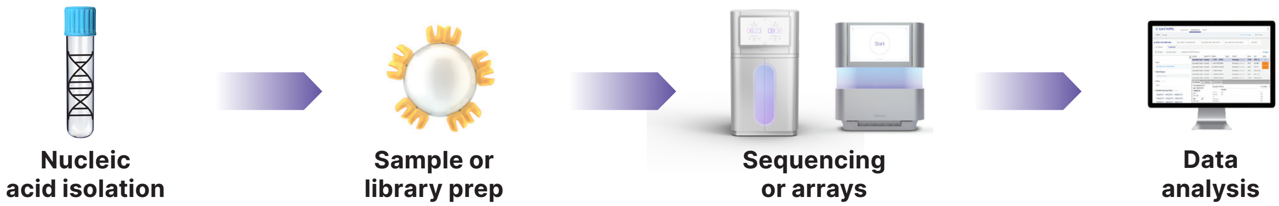


Figure 4: General multiomics workflow.

Library/sample preparation

The method used for library preparation will depend on the experimental question. We point out specific sample/library prep kits for each multiomics method in Section 5. Illumina library prep kits are optimized specifically for Illumina sequencing systems as well as the secondary analysis platform, DRAGEN™ Bio-IT.

Box 1: Illumina library prep reagents.

Illumina library prep products include:

- Illumina DNA PCR-Free Prep (whole-genome sequencing)
- Illumina DNA Prep (whole-genome sequencing)
- Illumina DNA Prep with Enrichment (whole-exome sequencing)
- Illumina Stranded Total RNA Prep (whole-transcriptome sequencing)
- Illumina Stranded mRNA Prep (mRNA sequencing)
- Illumina RNA Prep with Enrichment (targeted RNA sequencing)
- Illumina Tagment DNA TDE1 Enzyme and Buffer Kits (ATAC-Seq)

Illumina sequencing

After library prep, sequencing is the next crucial step. Illumina uses sequencing by synthesis (SBS) chemistry that detects single bases as they are incorporated into growing DNA strands. Libraries are loaded onto flow cells and ran on Illumina sequencers.

Depending on the research question and the scale of the study, researchers may use the NovaSeq™ X series, NovaSeq 6000 system, or NextSeq™ 1000, or NextSeq 2000 system for sequencing.

The NovaSeq X and NovaSeq X Plus Sequencing Systems deliver extraordinary throughput and accuracy to perform data-intensive applications at production scale. Up to 16 Tb output (or 52 billion reads) support sequencing of more than 128 human genomes, ~1500 exomes, or over 1000 transcriptomes per dual flow cell run.

The NovaSeq 6000 system efficiently supports high-throughput whole-genome sequencing (up to 6 Tb per run), whole-exome sequencing (up to 500 exomes per run) and whole-transcriptome sequencing (up to 400 transcriptomes per run).

The NextSeq 1000 and NextSeq 2000 systems are benchtop instruments that support emerging and mid-throughput sequencing applications including targeted exome sequencing, single-cell profiling, chromatin analysis, and transcriptome sequencing.

Illumina methylation arrays

DNA methylation plays an essential role in regulating gene expression and facilitates responses to environmental stimuli. Changes in the DNA “methylome” and its impact on gene expression have been implicated in many biological processes and diseases. Methylation arrays enable high-throughput, cost-effective quantitative investigation of methylation sites across the genome.

Illumina Infinium MethylationEPIC Kit interrogates 850,000 CpG sites across the human epigenome and can be used to investigate genetic changes in various biological and regulatory processes. The Infinium Mouse Methylation BeadChip provides comprehensive coverage of the mouse methylome by analyzing 285,000 CpG sites per sample at single-nucleotide resolution. Illumina Infinium Custom Methylation BeadChip can be designed with between 3000 and 100,000 markers for a variety of applications. The high-throughput 24-sample Beadchip reduces per--sample and processing costs while maintaining the same robust methylation measurements provided by the Infinium assay.

FFPE or fresh/frozen tissue are processed with bisulfite conversion using the Infinium workflow. Samples are subsequently scanned on Illumina iScan™ System.

Methylation array workflows follow a similar format to what is shown in Figure 4, beginning with nucleic acid isolation, sample preparation, microarray analysis, and data analysis.

Data analysis

After sequencing is complete, data analysis can be performed. In general, the multiomics data analysis pipeline consists of three phases: primary analysis (also referred to as base calling), secondary analysis, and tertiary analysis. Several approaches and software are available for each analysis step. The approach you use will depend on the research objective and the omes you're studying.

Key multiomics data analysis terms are defined below.

Alignment

In this secondary data analysis step, researchers align sequences with reference sequences for comparison.

Multiplexing

To increase throughput of sequencing systems, large numbers of libraries with unique indexes can be pooled together, loaded into one lane of a sequencing flow cell, and sequenced in the same run.

Demultiplexing

This is a step in the analysis process where the researcher uses barcode information to separate, and identify which sequences come from which cells when multiple single cells are sequenced together.

Primary analysis

As discussed previously, Illumina SBS chemistry detects single bases as they are incorporated into growing DNA strands. Primary analysis is the first step in converting those base-associated signals into useful data by determining if they correspond to an A, T, G, or C. This step is completed automatically on Illumina sequencers, which generate raw data files in binary base call (BCL) format. In later steps, BCL files will be converted into other file types as researchers process and make sense of data.

Secondary analysis: Powered by Illumina DRAGEN BioIT Platform

Secondary analysis is where researchers begin to make sense of raw sequencing data. Illumina DRAGEN BioIT Platform features tools for every step of most secondary analysis pipelines, and is easily accessible on select Illumina sequencers (eg, NextSeq 1000 and NextSeq 2000 systems and NovaSeq X series), as a standalone physical server, or through cloud platforms like BaseSpace™ Sequence Hub or Illumina Connected Analytics (ICA). BaseSpace Sequence Hub is a great tool for researchers with or without bioinformatics experience to analyze data through graphic user interface (GUI)-based apps, while ICA is more suited for users with bioinformatics experience who perform high-volume data analysis. Regardless of where you access DRAGEN analysis, the platform provides accurate, comprehensive, and efficient data analysis solutions for a wide variety of NGS workflows.

The BCL sequence file format generated during primary analysis requires conversion to FASTQ format for use with Illumina, user-developed, or third-party secondary analysis tools. This can be accomplished through BCL Convert that can be accessed via DRAGEN on-board NextSeq 1000, NextSeq 2000 systems, or NovaSeq X series, on-premise server or through BaseSpace Sequence Hub and ICA cloud platforms.

DRAGEN secondary analysis includes an optional step for data compression using Original Read Archive (ORA), a lossless compression technology for FASTQ files. DRAGEN ORA can reduce FASTQ file sizes by up to five times, allowing for more manageable data storage.

Sequencing experiments may also be **multiplexed**, meaning that libraries from different samples were pooled and sequenced simultaneously in a single run. Running multiplexed sequencing experiments enables you to run many distinct samples in a single lane of a flow cell. This increases the sequencer's output and provides you with more data.

Each sample has a unique barcode that can be used to identify it post-sequencing. After conversion of BCL files to the standard FASTQ format, these multiplexed samples may undergo a necessary **demultiplexing** step. Demultiplexing separates and identifies sequences based on their unique barcodes, and assigns them to the cell or sample they came from. If samples were not multiplexed, then the demultiplexing step does not occur, because all reads are from a single sample.[30] On the NovaSeq X and NovaSeq X Plus systems, demultiplexing happens onboard in real time.

Subsequent steps in secondary analysis pipelines often involve aligning data to a reference genome (**alignment**). Researchers can then identify experimental variables such as sequence variants or differential gene expression between samples. As integrated into the NovaSeq X series, the DRAGEN platform can run multiple secondary analysis pipelines in parallel. Perform up to four simultaneous applications per flow cell in a single run. See Table 2 below for examples of DRAGEN secondary analysis tools that are available on BaseSpace Sequence Hub and ICA, and which are relevant to the methods we discuss in this guide.

Table 2: Illumina-developed multiomics data analysis solutions.

Software	Description/Application
DRAGEN Germline and Somatic	The DRAGEN Germline and Somatic apps provide end-to-end NGS analysis and is best for whole-exome and whole-genome sequencing data.
DRAGEN Enrichment	The DRAGEN Enrichment app provides rapid, end-to-end NGS analysis, including advanced error model calibration for increased accuracy for enrichment and hybridization-based exome panels.
DRAGEN RNA	The DRAGEN RNA pipeline includes an RNA-Seq (splicing-aware) aligner, as well as RNA-specific analysis components for gene expression quantification, gene fusion detection variant calling, and forced genotyping.
DRAGEN Differential Expression	DRAGEN Differential Expression application performs secondary analysis of RNA transcripts. It runs the DESeq2 algorithm on RNA quantification files produced by DRAGEN RNA app, to output genes and transcripts that are differentially expressed between two sample groups.
DRAGEN Original Read Archive (ORA) Compression	DRAGEN ORA is a lossless compression technology for multiomics data. DRAGEN ORA can reduce FASTQ file sizes by up to 5 times, allowing for more manageable data storage. This software also ensures no data in FASTQ files are lost, including read order.

Tertiary analysis: Data visualization and interpretation

This is the phase in the data analysis pipeline where you can visualize and interpret data.

Illumina BaseSpace Correlation Engine app is a popular example of a user-friendly tool that is often used for multiomic datasets.[31]

BaseSpace Correlation Engine is an interactive data analysis tool that can help validate results and test new hypothesis by enabling comparison of new sequencing data to 23,000 (and growing) scientific studies and public datasets.

Key BaseSpace Correlation Engine features include:

- A continually growing library of curated, public genomic data
- The ability to identify mechanisms of disease, drug targets, and prognostic or predictive biomarkers
- "Search Literature" and "Quick View" functions that can be used free of charge and with no login
- Access to an arsenal of web-based tools to mine data and create billions of novel correlations
- The opportunity to view pathways that play a role in disease development across multiple studies and data types
- Gene analysis functions that cut across more than 20,000 genomic studies to understand gene activity across different species
- The ability to analyze candidate molecules for pharmacokinetic and toxicity profiles
- Researchers can compare human data to model organism experimental results to gain context

For more information on BaseSpace Correlation Engine, check out the Correlation Engine Webpage. There are several free and commercially available software programs researchers can use for tertiary analysis.

See the tables below for a summary of free and commercial programs available.

Table 3: Open-source tertiary analysis software.

Software	Provider	Description
Seurat	Satija Lab satijalab.org/seurat	Seurat is an R package designed for QC, analysis and exploration of single-cell RNA-Seq data. Seurat enables users to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data.
t-SNE	Van der Maaten Lab lvdmaaten.github.io/tsne	t-distributed stochastic neighbor embedding (t-SNE) is a computational technique that visualizes high dimensional data by giving each data point a location in a two or three-dimensional map. t-SNE is commonly used to visualize subpopulations with single-cell sequencing data.
UMAP	Git Hub https://github.com/lmcinnes/umap	Uniform manifold approximation and projection (UMAP) is an algorithm for analysis of high dimensional data. It is an alternative to t-SNE offering faster computation times.
Monocle 3	Trapnell Lab https://cole-trapnell-lab.github.io/monocle3/	Monocle is an R-based single-cell RNA-Seq analysis software designed to determine cell development trajectory. Monocle is ideal for experiments where there are known beginning and terminal cell states.
Human Cell Atlas	Broad Institute Humancellatlas.org	The Human Cell Atlas is a consortium effort that will curate a data coordination platform intended to provide four key components: intake services for data submission, synchronized data storage across multiple clouds, standardized secondary analysis pipelines and portals for data access, tertiary analysis and visualization.
Gene Set Enrichment Analysis	UC San Diego/Broad Institute http://www.gsea-msigdb.org/gsea/index.jsp	Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant and concordant differences between two biological states.

Table 4: Commercially available tertiary analysis software.

Software	Provider	Description
SeqGeq™	BD Biosciences	SeqGeq is a desktop application with an easy-to-use interface for advanced data analysis, exploration, and visualization of single-cell gene expression data. SeqGeq offers powerful data reduction and population identification tools.
Partek Flow	Partek	Partek Flow is a software analysis solution for NGS data applications. It provides robust statistical algorithms, information-rich visualizations, and cutting-edge genomic tools, enabling researchers of all skill levels to confidently perform data analysis.
CytoBank	CytoBank/Beckman Coulter	CytoBank is a cloud-based platform designed for analysis and visualization of multiple single-cell data sets simultaneously.
Loupe Cell Browser	10x Genomics	The Loupe Cell Browser is designed to enable users to quickly and interactively find significant genes, cell types, and substructure within single-cell data.
Tapestri Insights	MissionBio	Tapestri Insight provides sequence import, data analysis and visualization for single-cell DNA analysis.

An extra step: Integrating your data with Illumina Connected Analytics (ICA)

The use of multiomics methods has allowed researchers to generate data at unprecedented levels. This large amount of data being generated far outpaces any organization's ability to extract relevant biological and clinical insights.

We built Illumina Connected Analytics (ICA) for researchers with bioinformatics expertise who are interested in analyzing data at scale. ICA is a comprehensive cloud-based data management and analysis platform that allows you to share, aggregate, and explore large volumes of sequencing data in a secure, scalable, and flexible environment.

ICA analysis pipelines are highly customizable and can also integrate DRAGEN analysis tools. ICA is also incredibly secure, making it a great option for clinical research institutions.

ICA offers:

- Direct integration with the data generation workflow using Illumina sequencing systems
- Powerful secondary analysis with the DRAGEN Bio-IT Platform
- Scalable data aggregation
- Secure data storage
- A dynamic and interactive data science environment for advanced machine learning and artificial intelligence

For more information on using ICA, read the Illumina Connected Analytics guide.

Array analysis

Researchers can analyze and visualize data from arrays with GenomeStudio™ Software. The GenomeStudio Methylation module supports the analysis of Infinium methylation array data. The module calculates methylation levels and analyzes differential methylation levels between experimental groups. It enables you to view CpG methylation status across the genome with the Illumina Genome Browser and Illumina Chromosome Browser.

Multiomics workflows

5.1: Genomics methods

Method 1: Whole-genome sequencing (WGS)

WGS analyzes the whole genome of a population of cells or of tissue samples. Using WGS, researchers can uncover genetic events that contribute to disease beyond protein-coding variants.

DNA extraction/library preparation

There are several DNA extraction methods that can be used to extract DNA for WGS from fresh frozen (FF) samples or from formalin fixed paraffin embedded (FFPE) samples. Once DNA is extracted, libraries can be prepared for Illumina sequencing using the kits in Table 5 below.

Table 5: Library prep kits for WGS.

Sample type	Illumina kit
Human	Illumina DNA PCR-Free Prep
Bacterial/any species	Illumina DNA Prep

Sequencing

Illumina recommends the NovaSeq X series or NovaSeq 6000 system for WGS.

Data analysis

We recommend using the DRAGEN platform for WGS data analysis. The DRAGEN platform can process data for an entire human genome at 30× coverage in about 25 minutes. You can access DRAGEN whole-genome pipelines (germline and somatic) on your sequencer or via the cloud-based BaseSpace Sequence Hub and Illumina Connected Analytics.[32] In BaseSpace Sequence Hub, you can monitor runs in real time while securely streaming data directly from the instruments into the ecosystem.

Workflow summary

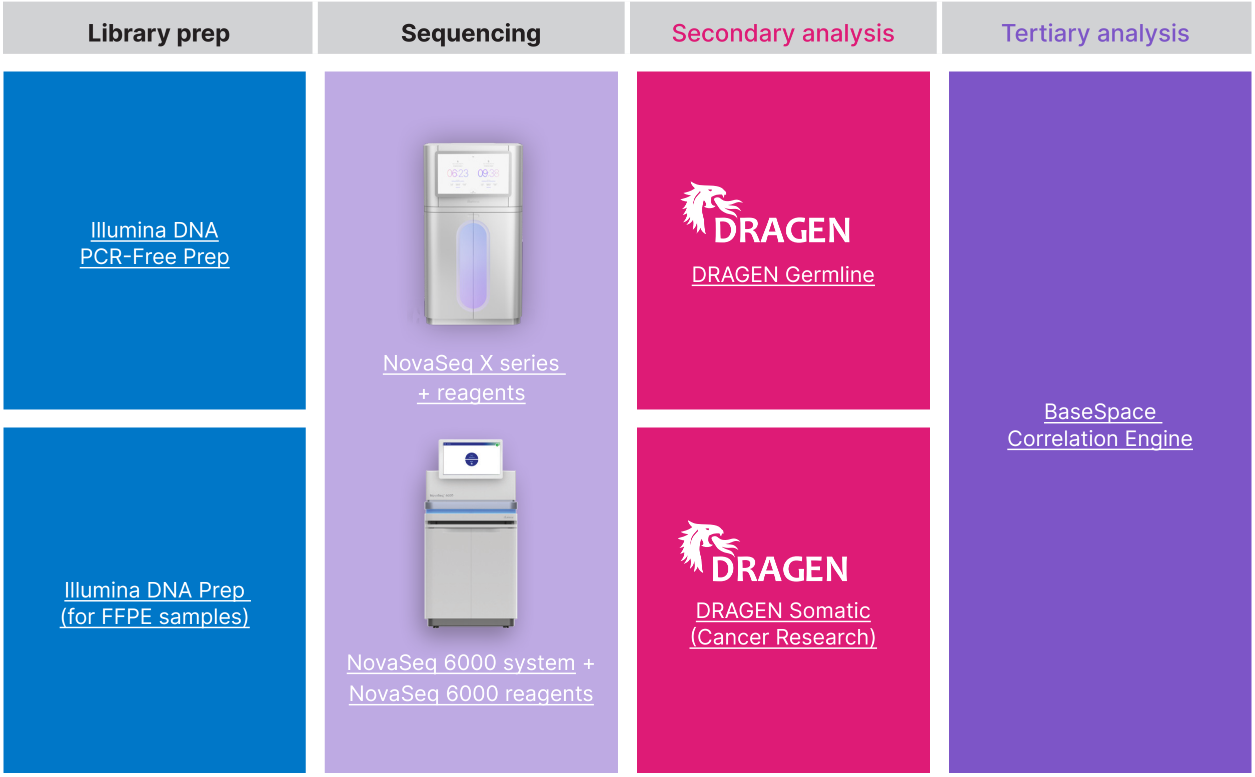

Library prep
Sequencing
Secondary analysis
Tertiary analysis
Illumina DNA PCR-Free Prep
Illumina DNA Prep (for FFPE samples)
NovaSeq X series + reagents
NovaSeq 6000 system + NovaSeq 6000 reagents
DRAGEN
DRAGEN Germline
DRAGEN
DRAGEN Somatic (Cancer Research)
BaseSpace Correlation Engine

Figure 5: Multiomics workflow summary for WGS.

Method 2: Whole-exome sequencing (WES)

WES allows researchers to analyze the portion of the genome responsible for coding proteins (the exome). While the exome represents less than 2% of the entire genome, it accounts for 85% of disease-related variants.[33] Using WES, researchers can study which protein-coding variants contribute to disease or dysfunction.

Advantages of exome sequencing:

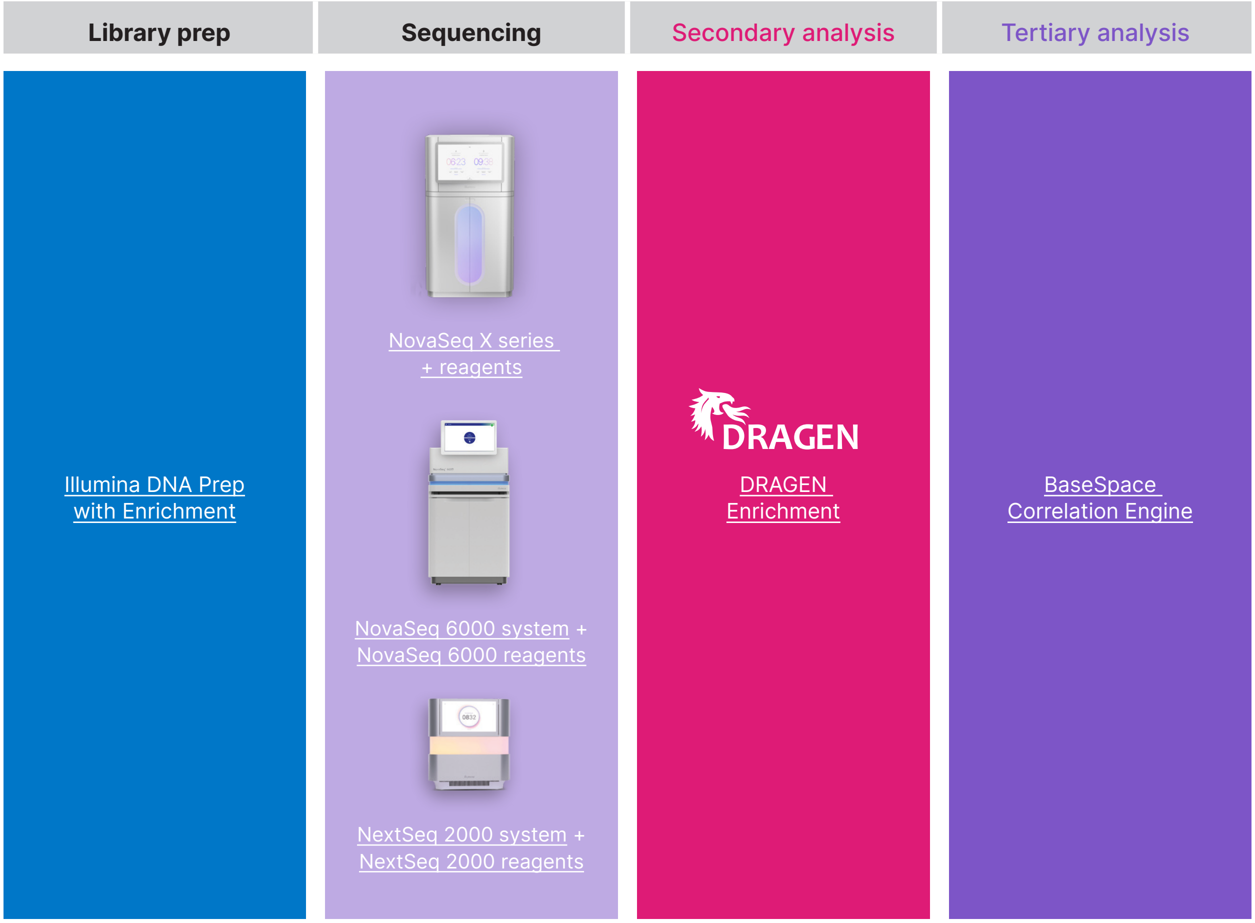
- WES identifies variants across a wide range of applications
- It achieves comprehensive coverage of coding regions of the genome
- WES is a cost-effective alternative to WGS (4–5 Gb of sequencing data versus up to 90 Gb of data for the whole human genome)
- Because of the above, the data generated from WES is more manageable and is easier to analyze

DNA extraction/library preparation
After DNA extraction, libraries for WES can be prepared using Illumina DNA Prep with Enrichment.

Sequencing
Illumina recommends the NextSeq 2000, NovaSeq 6000 systems, or NovaSeq X series for WES.

Data analysis
We recommend using the DRAGEN platform either on BaseSpace Sequence Hub or on a DRAGEN server to obtain data from WES. In BaseSpace Sequence Hub, you can monitor runs in real time while securely streaming data directly from the instruments into the ecosystem.

Workflow summary



Library prep	Sequencing	Secondary analysis	Tertiary analysis
Illumina DNA Prep with Enrichment	NovaSeq X series + reagents NovaSeq 6000 system + NovaSeq 6000 reagents NextSeq 2000 system + NextSeq 2000 reagents	DRAGEN Enrichment	BaseSpace Correlation Engine

Figure 6: Workflow summary for WES.

5.2: Transcriptomics methods

Method 3: Whole-transcriptome sequencing

Coding and noncoding forms of RNA play crucial roles in cellular regulation and disease pathogenesis. Whole-transcriptome sequencing or total RNA sequencing (total RNA-Seq) analyzes and detects coding and non-coding forms of RNA, such as microRNA, to provide a comprehensive view of the transcriptome. Total RNA-Seq can accurately measure known transcript abundance as well as identify novel variants.

Advantages of total RNA-Seq include:

- It captures both known sequences and novel variants
- Total RNA-Seq allows researchers to identify biomarkers across a broad range of transcripts
- It provides comprehensive insights into phenotypes of interest

Library preparation

Illumina recommends the Illumina Stranded Total RNA Prep with Ribo-Zero Plus for rapid library preparation for total RNA-Seq from a broad range of sample types. This enhanced library prep kit allows you to remove rRNA so you can focus on analyzing the actual transcriptome.

Sequencing

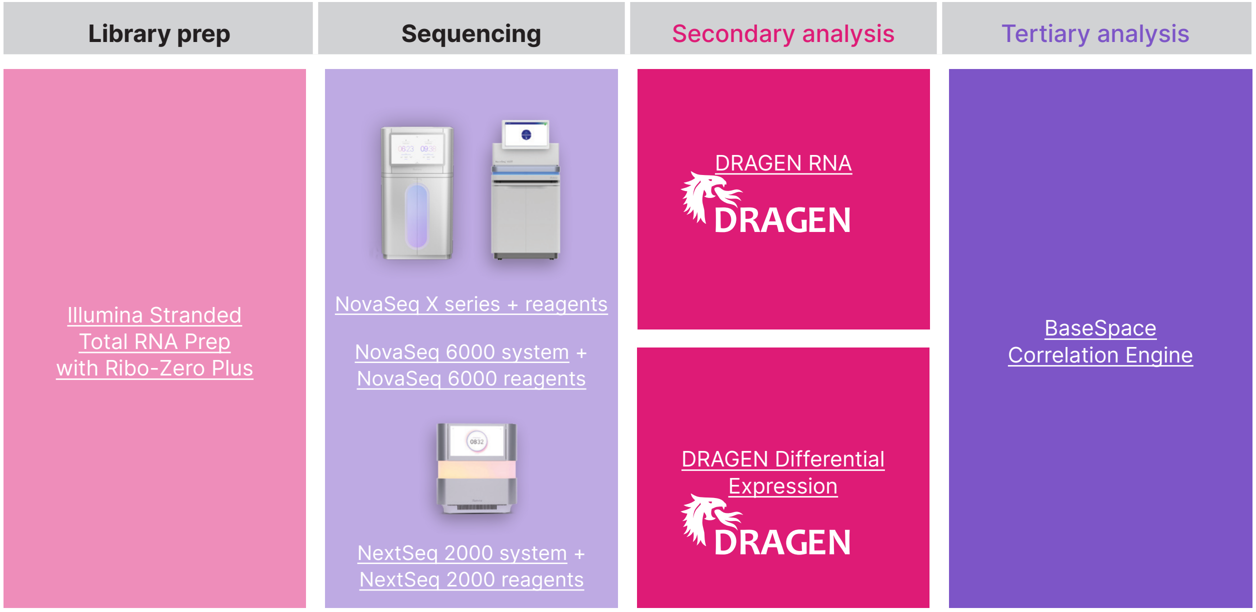
Table 6: Recommended sequencing systems for total RNA-Seq.

Sequencing System	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

Data analysis

Researchers can perform secondary analysis of total RNA-Seq data using the DRAGEN RNA pipeline or Differential Expression apps to obtain differential expression results at the gene and transcript levels. Tertiary analysis can be performed using the BaseSpace Sequence Hub or the BaseSpace Correlation Engine.

Workflow summary



Library prep	Sequencing	Secondary analysis	Tertiary analysis
Illumina Stranded Total RNA Prep with Ribo-Zero Plus	NovaSeq X series + reagents NovaSeq 6000 system + NovaSeq 6000 reagents NextSeq 2000 system + NextSeq 2000 reagents	DRAGEN RNA DRAGEN Differential Expression	BaseSpace Correlation Engine

Figure 7: Workflow summary for total RNA-Seq.

Method 4: mRNA sequencing

Messenger RNA accounts for around 2% of the whole-transcriptome and is composed of polyA tailed RNA that codes for proteins. mRNA sequencing (mRNA-Seq) provides an unbiased and complete view of the coding transcriptome.

Advantages of mRNA-Seq include:

- Compared to total RNA-Seq, mRNA-Seq allows researchers to focus on the coding transcriptome which also means less but targeted data
- mRNA-Seq captures both known and novel features
- It offers a broad dynamic range and therefore enables more sensitive and accurate measurement of fold changes in gene expression

Library preparation

Illumina Stranded mRNA Prep is a simple, scalable rapid library preparation solution for analyzing the coding transcriptome with as little as 25 ng RNA input. For FFPE samples, we recommend Illumina RNA Prep with Enrichment.

Sequencing

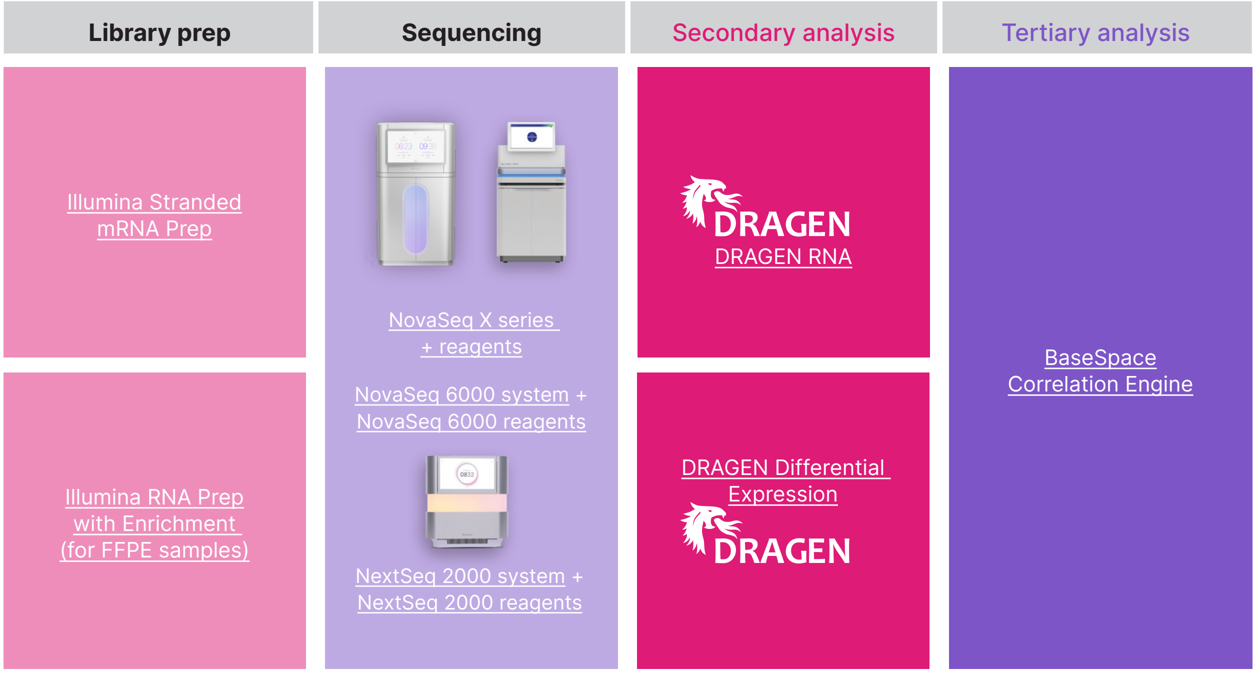
Table 7: Recommended sequencers for mRNA-Seq.

Sequencing System	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

Data analysis

Researchers can perform secondary analysis of mRNA-Seq data using the DRAGEN RNA pipeline or Differential Expression apps to obtain differential expression results at the gene and transcript levels. Tertiary analysis can be performed using the BaseSpace Sequence Hub or the BaseSpace Correlation Engine.

Workflow summary



Library prep

Illumina Stranded mRNA Prep

Illumina RNA Prep with Enrichment (for FFPE samples)

Sequencing

NovaSeq X series + reagents

NovaSeq 6000 system + NovaSeq 6000 reagents

NextSeq 2000 system + NextSeq 2000 reagents

Secondary analysis

DRAGEN RNA

DRAGEN Differential Expression

Tertiary analysis

BaseSpace Correlation Engine

Figure 8: Workflow summary for mRNA-Seq.

5.3: Proteomics methods (in combination with other omics)

Method 5: Cellular indexing of transcriptomes and epitopes by sequencing (CITE-Seq)

CITE-Seq uses oligonucleotide-labeled antibodies to measure proteins and RNA in the same experiment. CITE-Seq is a high-throughput multiomics tool that allows researchers to study protein expression and the intricacies of the cellular transcriptome both at the single-cell level and for spatial analysis (see Method 7).[34]

The advantages of using CITE-Seq include:

- This combined proteomics/transcriptomics approach allows researchers to tie RNA expression directly to a phenotype
- Since CITE-Seq method studies at two omes at once, the workflow is shorter than combining two individual omics methods

Library preparation

Whether you plan on doing single-cell analysis or spatial analysis, there are kits to support your experimental path. Several third-party products can be used to prepare libraries that are compatible with Illumina sequencers (Table 8).

Table 8: Library prepation kits for CITE-Seq.

Single-cell analysis library prep kits
BioLegend TotalSeq A, B, C Reagents
BD® AbSeq Assay
10x Single Cell Profiling
10x Single Cell Gene Expression

Sequencing

Table 9: Recommended sequencers for CITE-Seq.

Sequencing System	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

All of these systems can be used for single-cell and spatial analysis.

Data analysis

The secondary or tertiary data analysis method you will use for CITE-Seq data will depend on the library prep method.

Table 10: Data analysis software for CITE-Seq based on the library preparation method.

Data analysis method	CITE-Seq library prep kit	Secondary/tertiary analysis
Single-cell analysis	BioLegend Total Seq-A, -B, -C Reagents	BioLegend Multiomics Analysis Software
	BD AbSeq Assay	BD SeqGeq Software
	10x Single Cell Profiling	10x Cell Ranger Software
	10x Single Cell Gene Expression	10x Cell Ranger Software

Workflow summary

Library prep	Sequencing	Secondary analysis	Tertiary analysis
TotalSeq – A, B or C Reagents	NovaSeq X series + reagents NovaSeq 6000 system + NovaSeq 6000 reagents NextSeq 2000 system + NextSeq 2000 reagents	BioLegend® Multiomics Analysis Software (MAS)	
BD AbSeq Assay		BD SeqGeq Software	
10x Single Cell Immune Profiling		Cell Ranger Software 10x GENOMICS	
10x Single Cell Gene Expression			

Figure 9: Workflow summary for CITE-Seq.

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

Method 6: Bulk epitope and nucleic acid sequencing (BEN-Seq)

BEN-Seq is another multiomics approach that studies both proteins and RNA in one experimental workflow. BEN-Seq uses oligonucleotide-linked antibodies to detect proteins within a sample. Following antibody binding, the oligonucleotides are used to prepare libraries that are consequently sequenced. BEN-Seq interrogates proteins and RNA in bulk-cell populations.

Advantages of using BEN-Seq include:

- BEN-Seq detects significantly more proteins (hundreds) more efficiently than traditional methods like Western blotting or ELISA
- BEN-Seq achieves similar levels of accuracy for measuring proteins as flow cytometry with the added benefit of measuring RNA expression on a high-throughput scale

Library preparation

Libraries for BEN-Seq can be prepared using:

- BioLegend TotalSeq A Reagents
- Illumina Stranded mRNA Prep
- Illumina RNA Prep with Enrichment (for FFPE samples)

Sequencing

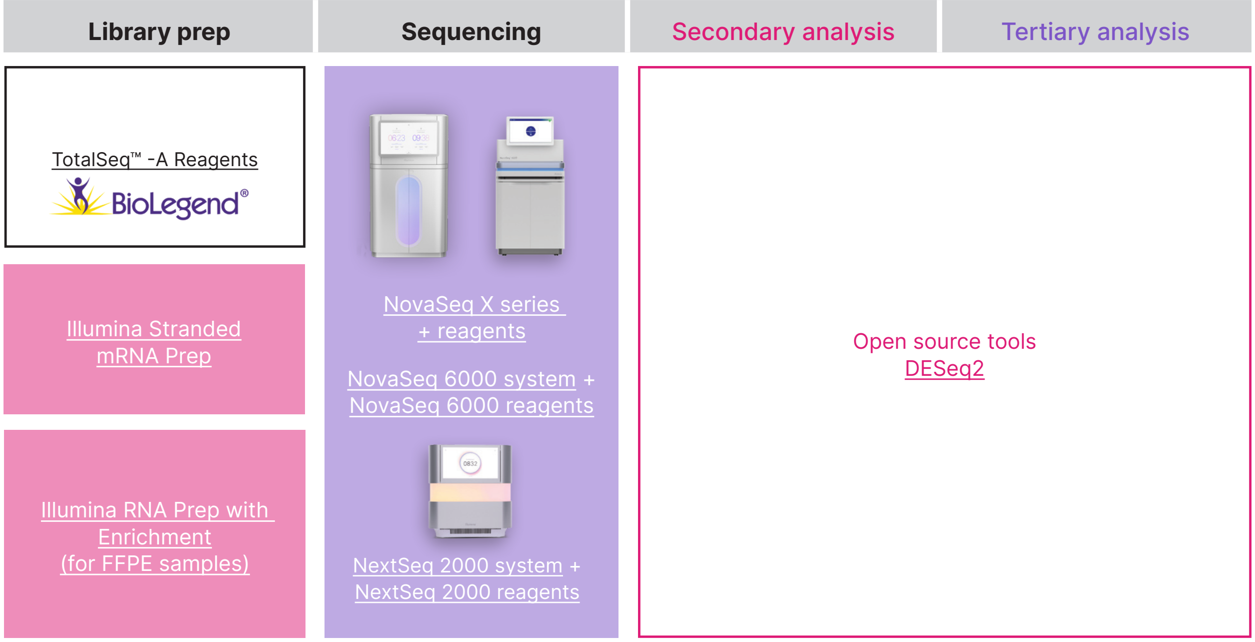
Table 11: Recommended sequencers for BEN-Seq.

Sequencing system	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

Data analysis

DESeq2 and several open-source tools can be used in BEN-Seq secondary and tertiary data analysis.

Workflow summary



Library prep

Sequencing

Secondary analysis

Tertiary analysis

TotalSeq™ -A Reagents BioLegend®

Illumina Stranded mRNA Prep

Illumina RNA Prep with Enrichment (for FFPE samples)

NovaSeq X series + reagents

NovaSeq 6000 system + NovaSeq 6000 reagents

NextSeq 2000 system + NextSeq 2000 reagents

Open source tools DESeq2

Figure 10: Multiomics workflow summary for BEN-Seq.

Method 7: Spatial transcriptomics and proteomics

Researchers can use spatial analysis to study proteins and RNA in an unperturbed tissue microenvironment. A microscope slide coated with spatially barcoded bead arrays captures the RNA from frozen histological tissue sections. The captured RNA undergoes reverse transcription into cDNA and is consequently sequenced. Spatial transcriptomics allows researchers to get as close to the realistic processes that happen in whole organismal systems, as possible.

Key benefits of using spatial transcriptomics over a method like RNA-Seq include:

- Spatial analysis conserves the spatial context for gene expression which removes the need for tissue dissociation
- RNA profiling can be combined with immunofluorescence or histochemical staining and imaging on the same sample
- Since this is one streamlined workflow, researchers can collect data in a shorter timeframe

Sample preparation

Samples used in spatial transcriptomics/proteomics experiments can be prepared using the following kits.

- 10x Genomics Visium Spatial Gene Expression
- NanoString GeoMx® RNA Assays
- NanoString GeoMx® Protein Assays

Sequencing

Table 12: Recommended sequencers for spatial analysis.

Sequencing system	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

Data analysis

The secondary or tertiary data analysis method you will use for data analysis will depend on the library prep method.

Table 13: Data analysis solutions for spatial analysis based on library prep kit.

Data analysis method	Library prep kit	Secondary/tertiary analysis
Spatial analysis	10x Genomics Visium Spatial Gene Expression	10x Space Ranger and 10x Loupe Browser
	NanoString GeoMx RNA Assays	BaseSpace Sequence Hub/ NanoString GeoMx Spatial Biology Data Analysis
	NanoString GeoMx Protein Assays	BaseSpace Sequence Hub/ NanoString GeoMx Spatial Biology Data Analysis

Workflow summary

Library prep	Sequencing	Secondary analysis	Tertiary analysis
Visium Spatial Gene Expression. GeoMx Digital Spatial Profiler GeoMx RNA Assays GeoMx Protein Assays	NovaSeq X series + reagents NovaSeq 6000 system + NovaSeq 6000 reagents NextSeq 2000 system + NextSeq 2000 reagents	10x GENOMICS Visium Spatial Gene Expression nanoString BaseSpace® SEQUENCE HUB GeoMx - Spatial Biology Data Analysis	

Figure 11: Workflow summary for spatial analysis.

Method 8: Proteogenomics

Researchers can use proteogenomics to resolve proteomic and genomic questions at the resolution of a single cell. Single cells are first stained with uniquely barcoded oligonucleotide-conjugated antibodies. Cells are then encapsulated so that proteins and DNA are released. The protein and DNA for each cell can be analyzed at this point.

The advantage of using proteogenomics is that researchers can directly link a genotype with a phenotype to fully understand biological processes and diseases.

Sample preparation
Samples for proteogenomics experiments can be prepared using BioLegend TotalSeq D Reagents and Mission Bio's Tapestri Single-Cell Platform.

Sequencing
Sequencing of proteogenomics libraries can be done on the NextSeq 2000 system.

Data analysis
Researchers can use the Mission Bio Tapestri Pipeline for secondary and tertiary analysis of proteogenomics data.

Workflow summary

Library prep	Sequencing	Secondary analysis	Tertiary analysis
BioLegend® TotalSeq - D Reagents	 NextSeq 2000 system + NextSeq 2000 reagents	mission bio Tapestri Pipeline Tapestri Insights	
mission bio Tapestri Platform Tapestri Single-Cell Panels			

Figure 12: Workflow summary for proteogenomics

5.4: Epigenomics methods

Method 9: Assay for transposase accessible chromatin with sequencing (ATAC-Seq)

ATAC-Seq is an epigenomic discovery tool for mapping chromatin accessibility across the genome. This approach analyzes DNA accessibility using the Tn5 transposase. The Tn5 transposase inserts sequencing adapters into open chromatin regions. Researchers can then use sequencing to locate regions of increased chromatin accessibility. ATAC-Seq allows researchers to study how these regions impact gene expression and can be used to study both single cells and cell populations (bulk-cell applications).[35]

ATAC-Seq is useful for studying:

- Nucleosome mapping
- Transcription factor binding
- Novel enhancers
- Regulatory mechanisms underlying disease
- Cell-specific regulatory mechanisms
- Evolution
- Novel biomarkers

Sample preparation

We recommend the Illumina Tagment DNA TDE1 Enzyme and Buffer Kits for preparing samples that will be used in ATAC-Seq.

Sequencing

Table 14: Recommended sequencers for ATAC-Seq.

Sequencing system	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

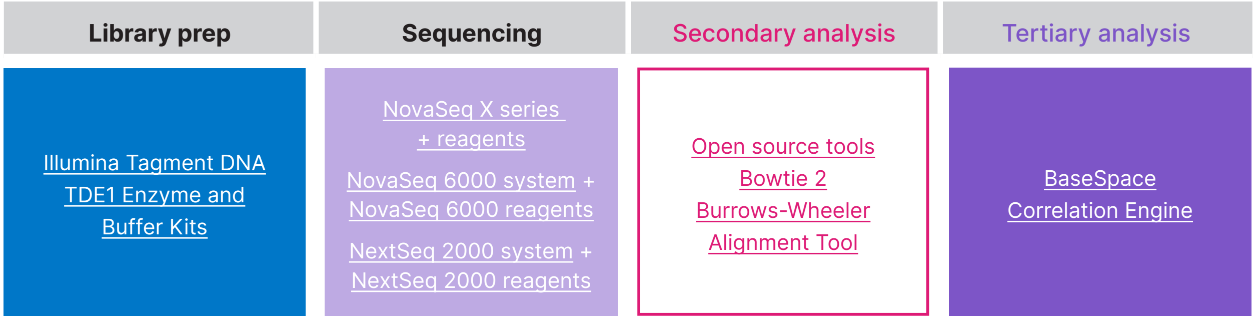
Data analysis

We recommend the following tools for analyzing ATAC-Seq data. For both primary and secondary analysis, you can choose one of the two tools listed for each.

Table 15: Data analysis solutions for ATAC-Seq.

Secondary analysis	Tertiary analysis (visualization)
Genrich	BaseSpace Correlation Engine
MACS2	
Burrows-Wheeler Alignment (BWA) Tool	
Bowtie 2	

Workflow summary



Library prep	Sequencing	Secondary analysis	Tertiary analysis
Illumina Tagment DNA TDE1 Enzyme and Buffer Kits	NovaSeq X series + reagents NovaSeq 6000 system + NovaSeq 6000 reagents NextSeq 2000 system + NextSeq 2000 reagents	Open source tools Bowtie 2 Burrows-Wheeler Alignment Tool	BaseSpace Correlation Engine

Figure 13: Workflow summary for ATAC-Seq.

Method 10: Methylation sequencing

DNA methylation plays a crucial role in regulating gene expression. Changes in DNA methylation patterns has an impact on many important human diseases including cancer, multiple sclerosis, diabetes, and addiction. Methylation sequencing therefore allows researchers to gain valuable insight into gene regulation and identify potential biomarkers.

Methylation sequencing can be used to:

- Discover methylation patterns of CpG, CHH, and CHG regions across the genome
- Study emerging regions of interest in the human genome identified by programs like ENCODE

Library preparation

Libraries for methylation sequencing can be prepared using the TruSeq Methyl Capture EPIC Library Prep Kit. This kit allows researchers to prepare enrichment-based bisulfite sequencing libraries from as little as 500 ng of human DNA samples. Methylation sequencing can be carried out on the NextSeq 2000 or NovaSeq 6000 systems, or the NovaSeq X series.

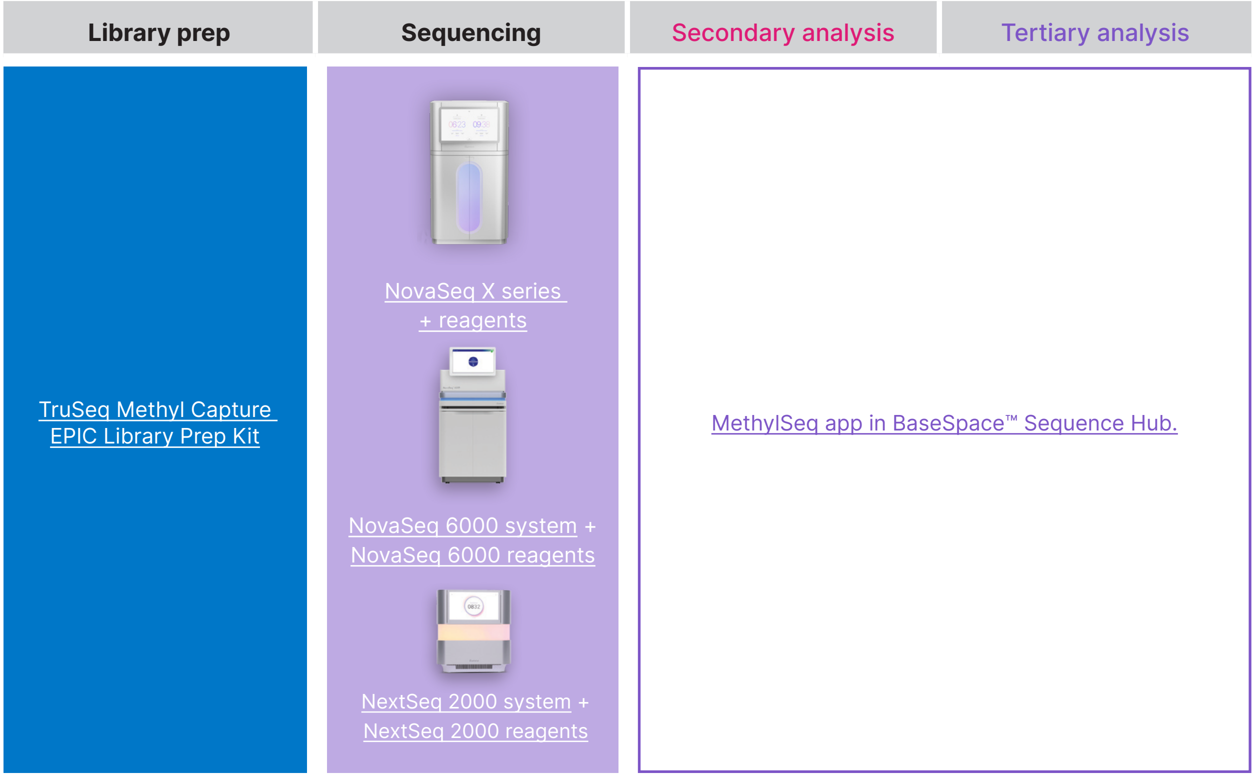
Table 16: Recommended sequencers for methylation sequencing

Sequencing system	Advantages
NovaSeq X/NovaSeq X Plus	Maximum throughput, operational simplicity, and lowest cost per sample of Illumina high-throughput systems
NovaSeq 6000	Robust, scalable high-throughput system
NextSeq 1000/NextSeq 2000	Flexible and intuitive mid-throughput benchtop system

Data analysis

The MethylSeq app in BaseSpace Sequence Hub can be used for secondary and tertiary analysis of methylation sequencing data.

Workflow summary



Library prep

Sequencing

Secondary analysis

Tertiary analysis

TruSeq Methyl Capture EPIC Library Prep Kit

NovaSeq X series + reagents

NovaSeq 6000 system + NovaSeq 6000 reagents

NextSeq 2000 system + NextSeq 2000 reagents

MethylSeq app in BaseSpace™ Sequence Hub.

Figure 14: Workflow summary for methylation sequencing.

Method 11: Methylation arrays

Methylation arrays enables scientists to quantitatively interrogate methylation sites across the epigenome. Microarray technology paved the way for current NGS technologies. Illumina methylation arrays depend on Infinium technology, which allows for the reliable measurement of methylation status with single-base resolution.

Methylation arrays are especially useful in:

- Developmental biology studies
- Transgenic mouse models of disease experiments
- Aging and epigenetic clock research
- Epigenome-wide association studies
- Biomarker discovery

Sample preparation

Illumina methylation arrays follow user-friendly and streamlined workflows that enable the processing of up to 96 samples with as little as 250 ng DNA. We recommend the Infinium MethylationEPIC BeadChip (over 850,000 markers) for robust methylation profiling in human samples for CpG islands, genes, and enhancers. The Infinium Mouse Methylation BeadChip is formulated for sampling murine samples. The Infinium Mouse Methylation BeadChip features over 285,000 markers.

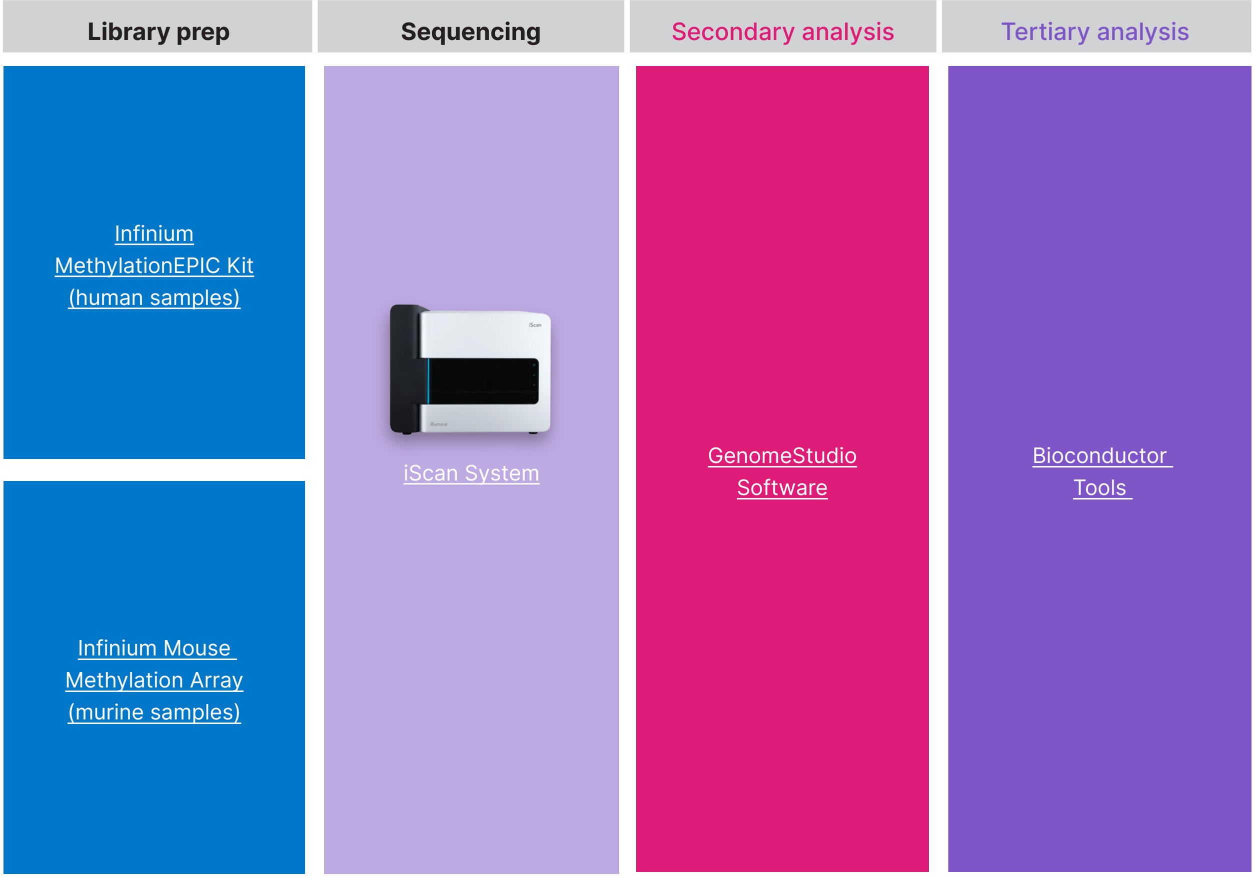
Array processing

The iScan System supports rapid and accurate high-throughput BeadChip processing.

Data analysis

The Methylation Module in GenomeStudio is routinely used for analyzing methylation array data. Tertiary analysis can be completed in the BaseSpace Correlation Engine.

Workflow summary



Library prep	Sequencing	Secondary analysis	Tertiary analysis
Infinium MethylationEPIC Kit (human samples) Infinium Mouse Methylation Array (murine samples)	iScan System	GenomeStudio Software	Bioconductor Tools

Figure 15: Workflow summary for methylation arrays.

References

[1]Anand P, Kunnumakkara AB, Sundaram C, et al. Cancer is a preventable disease that requires major lifestyle changes. *Pharm Res*. 2008;25(9):2097-2116.

[2]Pon JR, Marra MA. Driver and passenger mutations in cancer. *Annu Rev Pathol*. 2015;10:25-50.

[3]Kumar S, Warrell J, Li S, et al. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell*. 2020;180(5):915-927.e16.

[4]Newell F, Pires da Silva I, Johansson PA, et al. Multiomic profiling of checkpoint inhibitor-treated melanoma: Identifying predictors of response and resistance, and markers of biological discordance. *Cancer Cell*. 2022;40(1):88-102.e7.

[5]Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. *Nature*. 2013;501(7467):328-337.

[6]Gahl WA, Markello TC, Toro C, et al. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med*. 2012;14(1):51-59.

[7]Abela L, Simmons L, Steindl K et al. 2016; N(8)-acetylspermidine as a potential plasma biomarker for Snyder-Robinson syndrome identified by clinical metabolomics. *J Inherit Metab Dis* 39:131–137

[8]Abela L, Spiegel R, Crowther LM et al. 2017; Plasma metabolomics reveals a diagnostic metabolic fingerprint for mitochondrial aconitase (ACO2) deficiency. *PLoS One* 12:e0176363

[9]Tarailo-Graovac M, Shyr C, Ross CJ et al. 2016; Exome sequencing and the Management of Neurometabolic Disorders. *N Engl J Med* 374:2246–2255

[10]Bick D, Jones M, Taylor SL, Taft RJ, Belmont J. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. *J Med Genet*. 2019;56(12):783-791.

[11]Krzyszczyk P, Acevedo A, Davidoff EJ, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology (Singap World Sci)*. 2018;6(3-4):79-100.

[12]Population matters: Biobanks accelerate geno-pheno discoveries. Illumina. 2020; *Nature advertorial*.

[13]Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell*. 2019;177(1):26-31.

[14]García-Dorival I, Wu W, Armstrong SD, et al. Elucidation of the Cellular Interactome of Ebola Virus Nucleoprotein and Identification of Therapeutic Targets. *Journal of Proteome Research*. 2016; 15 (12): 4290-4303.

[15]Scaturro P, Kastner AL, Pichlmair A. Chasing Intracellular Zika Virus Using Proteomics. *Viruses*. 2019; 11 (9): 878.

[16]Coombs KM, Berard A, Xu W, et al. Quantitative Proteomic Analyses of Influenza Virus-Infected Cultured Human Lung Cells. *J Virol*. 2010; 84(20): 10888-10906.

[17]Aggarwal S, Acharjee A, Mukherjee A, Baker MS, Srivastava S. Role of Multiomics Data to Understand Host-Pathogen Interactions in COVID-19 Pathogenesis. *J Proteome Res*. 2021;20(2):1107-1132.

[18]Wu F, Zhao S, Yu B, et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature*. 2020; 579(7798): 265-269.

[19]Zhou P, Yang XL, Wang XG, et al. A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature*. 2020; 579(7798): 270-273.

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

[20] Corman VM, Landt, O, Kaiser M, et al. Detection of 2019 Novel Coronavirus (2019-nCoV) by Real-Time RT-PCR. *Eurosurveillance*. 2020; 25(3): 2000045.

[21] Korber B, Fischer WM, Gnanakaran S, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus. *Cell*. 2020; 182(4): 812-827 (e19).

[22] Breijyeh Z, Karaman R. Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules*. 2020;25(24):5789.

[23] Morabito S, Miyoshi E, Michael N, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet*. 2021;53(8):1143-1155.

[24] Cardiovascular Disease. https://www.who.int/health-topics/cardiovascular-diseases. Accessed February 2022.

[25] Liu B, Pjanic M, Wang T, et al. Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. *Am J Hum Genet*. 2018;103(3):377-388.

[26] Nassiri, Liu, Patil, et al. A clinically applicable integrative molecular classification of meningiomas. *Nature*. 2021;597 119–125.

[27] Arunachalam PS, Wimmers F, Mok CKP, et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science*. 2020;369(6508):1210-1220.

[28] Cancer Stat Facts: Melanoma of the Skin. https://seer.cancer.gov/statfacts/html/melan.html. Accessed February 2022.

[29] Ranzoni AM, Tangherloni A, Berest I, et al. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell*. 2021;28(3):472-487.e7.

[30] Illumina. bcl2fastq Conversion User Guide. Published 2013. Accessed February 2022.

[31] Alnafakh R, Saretzki G, Midgley A, et al. Aberrant Dyskerin Expression Is Related to Proliferation and Poor Survival in Endometrial Cancer. *Cancers (Basel)*. 2021;13(2):273.

[32] Illumina. DRAGEN Bio-IT Platform. https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html. Accessed February 2022.

[33] Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet*. 2014;59(1):5-15.

[34] Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14(9):865-868.

[35] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29.9.

The time to get started is now

Here's why you should partner with Illumina.

Our multiomics reagents, equipment, and protocols have been validated with hundreds of publications in high-impact, peer-reviewed journals. (pubmed.ncbi.nlm.nih.gov/?term=illumina+sequencing)

Illumina technology has the broadest range of applications to enable multiomic analyses on one device and is built to support emerging changes in the multiomics landscape.

Illumina supports over 10,000 labs across 115 countries. 24/7 technical support professionals are ready to help you with any project.

Learn more about the power of multiomics

Visit our website to learn more about Illumina multiomics:

https://www.illumina.com/morewithmultiomics

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-AMR-00530 v2.0