illumina®

Analysis of Gene Expression and Regulation Studies: Critical Considerations

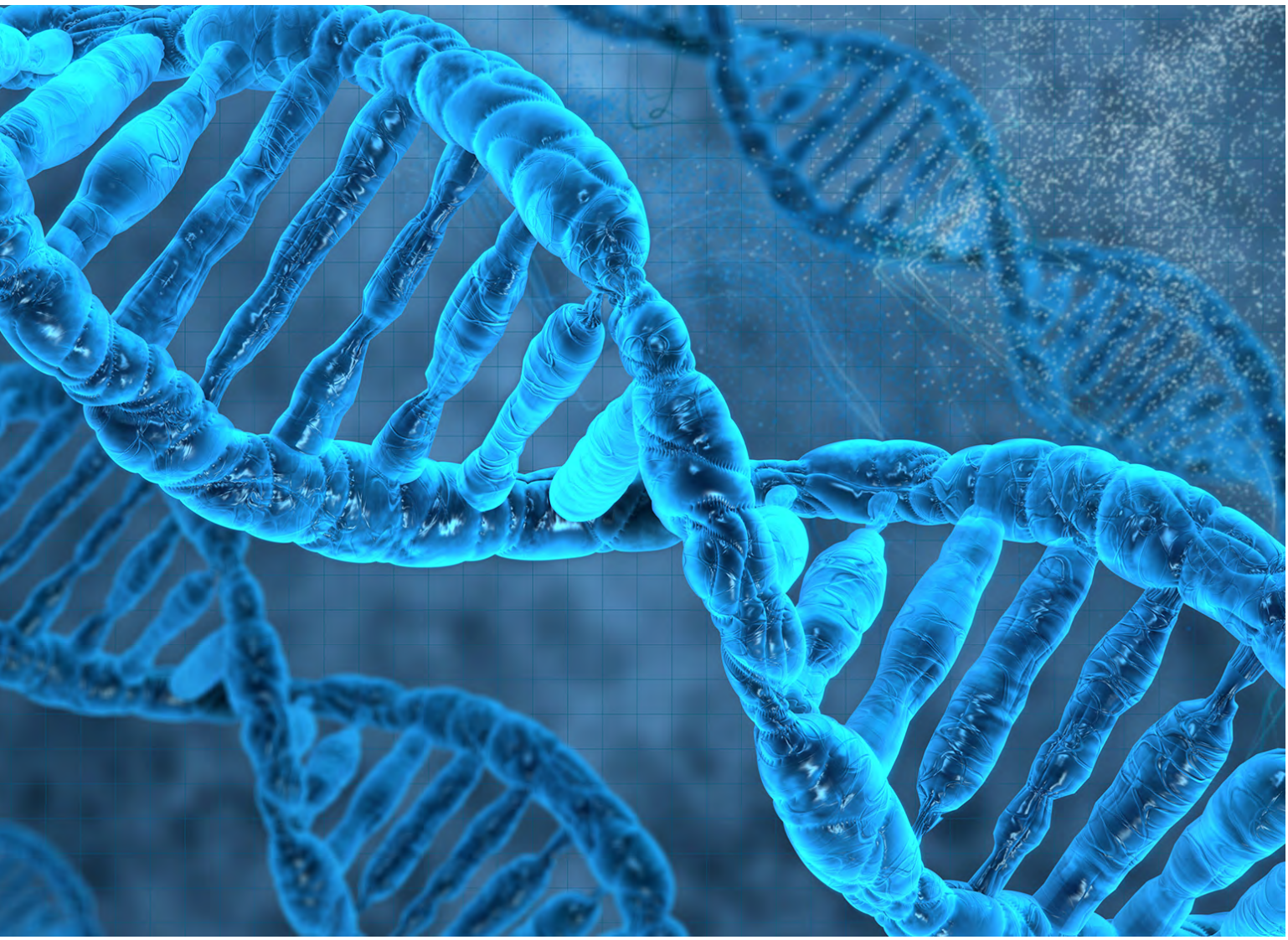
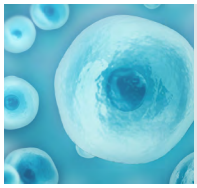
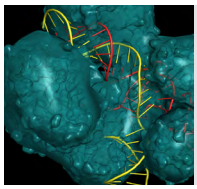


Table of contents

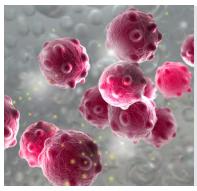
Introduction



How is gene expression controlled?



How are gene expression and protein production regulated by different types of cells?



How does the dysregulation of gene expression contribute to various disease states?

The answers to these questions and more lie in the study of gene expression.

Gene expression is the complex process by which DNA is translated into functional, biologically active units. In addition to encoding for proteins, genes encode for a vast array of additional nonprotein-coding RNA elements that drive essential biological processes, such as development and differentiation and, when dysregulated, the potential development of complex diseases.[1,2]

Next-generation sequencing (NGS)-based RNA sequencing (RNA-Seq) is a technological advance enabling scientists to push beyond the limits of traditional research methods. RNA-Seq is a highly sensitive and accurate tool that delivers a high-resolution, base-by-base view of coding and noncoding RNA activity for measuring gene expression across the transcriptome. This method can help elucidate previously undetected changes occurring in disease states, in response to therapeutics, under different environmental conditions, and across a broad range of other study designs.

To understand gene regulation and protein expression, advances in NGS methods have resulted in the development of the Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-Seq), Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq), and other methods, in addition to methylation sequencing and chromatin immunoprecipitation sequencing (ChIP-Seq).

Advances in cell isolation and sequencing technologies have resulted in the development and increased use of single-cell sequencing. This NGS method examines the genomes, epigenomes, transcriptomes, or proteomes of individual cells, providing a high-resolution view of cell-to-cell variation to reveal the cellular heterogeneity that drives complex biological systems.

Whether conducting bulk or single-cell studies, data analysis, visualization, and interpretation are critical steps for drawing insightful conclusions. With increasing options for investigating gene expression and regulation, there has been a remarkable diversification of bioinformatics pipelines, each with inherent strengths and weaknesses. Careful pipeline design and optimization throughout the analysis workflow is crucial for success.

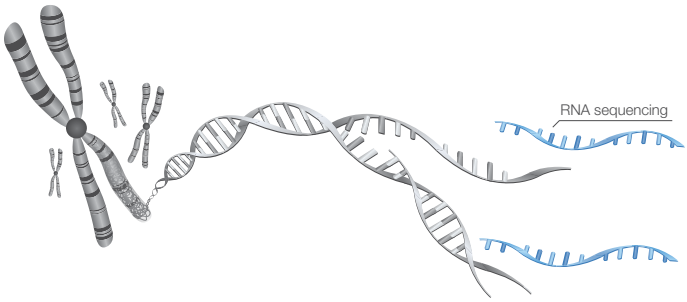
This e-book outlines every step of the NGS workflow, with a focus on the computational analysis pipelines available, for analyzing bulk and single-cell data. Important considerations and potential challenges are discussed, commercial offerings are presented, and advice is given for designing and executing a successful NGS-based gene expression and regulation study.

Chapter 1: Experimental considerations for NGS-based gene expression and regulation studies

There are various bulk sequencing methods available for investigating gene expression and regulation to understand normal cell development and disease mechanisms. For a higher resolution view, researchers can employ single-cell sequencing methods to elucidate the contribution of individual cells and overall function of complex tissues.

Bulk gene expression solutions

Illumina RNA library preparation



Advances in the Illumina portfolio of RNA library preparation kits deliver the high-quality data researchers require, with a streamlined workflow that can be completed within one standard working shift. Illumina offers three RNA library prep kits (Table 1):

- **Illumina Stranded Total RNA Prep** enables whole-transcriptome analysis using Ribo-Zero™ Plus, capturing coding and multiple forms of noncoding RNA to obtain a comprehensive picture of biology.
- **Illumina Stranded mRNA Prep** provides a cost-efficient option for coding RNA-focused analysis.
- **Illumina RNA Prep with Enrichment** brings bead-linked transposome (BLT) technology to RNA-Seq and provides a fast single-day RNA enrichment workflow for insights across many genomic positions.

Table 1: Illumina RNA library preparation solutions

	Illumina Stranded Total RNA Prep with Ribo-Zero Plus	Illumina Stranded mRNA Prep	Illumina RNA Prep with Enrichment
Description	For whole-transcriptome sequencing studies of multiple sample types. Includes Ribo-Zero Plus module for single-tube depletion of abundant transcripts from multiple species	A simple, cost-effective solution for analyzing the coding transcriptome with precise strand information	A reproducible, economical solution enabling targeted transcript detection and discovery from a broad range of sample types and inputs, including formalin-fixed, paraffin-embedded (FFPE) tissues and other low-quality samples
RNA analyzed	Captures coding RNA plus multiple forms of noncoding RNA	Captures the coding transcriptome with strand information	Captures the coding transcriptome when used with the Illumina Exome Panel
Method	whole-transcriptome sequencing	mRNA sequencing	mRNA sequencing, target enrichment
Assay time	~7 hours	6.5 hours	< 9 hours
Hands-on time	< 3 hours	< 3 hours	< 2 hours
FFPE support	Yes	No	Yes
Automation capability?	Yes	Yes	Yes
Input quantity	1-1000 ng standard-quality total RNA; 10 ng total RNA minimum recommended for optimal performance and FFPE samples	25-1000 ng standard-quality total RNA	10 ng total RNA from fresh/frozen samples, or 20 ng total RNA from FFPE samples
Mechanism of action	Enzymatic rRNA depletion, ligation-based addition of adapters and indexes	PolyA capture, ligation-based addition of adapters and indexes	Tagmentation with bead-linked transposomes
Link	Learn more	Learn more	Learn more

AmpliSeq™ for Illumina targeted panels

AmpliSeq for Illumina is a suite of AmpliSeq chemistry products that are compatible with Illumina NGS platforms. The AmpliSeq for Illumina solution offers a highly multiplexed PCR-based workflow for use with targets ranging from a few to hundreds in a single run. Combined with Illumina NGS technology, AmpliSeq for Illumina offers high-confidence data to researchers in a wide variety of application areas. AmpliSeq for Illumina works with RNA samples and requires as little as 1 ng of input. Users can select from predesigned panels or customizable content for various RNA-based applications. (Table 2):

- **AmpliSeq for Illumina Transcriptome Human Gene Expression Panel** to measure expression levels of > 20,000 human RefSeq genes.
- **AmpliSeq for Illumina Immune Response Panel** to investigate expression of 395 genes involved in tumor-immune system interactions.
- **AmpliSeq for Illumina Immune Repertoire Plus, TCR beta Panel** to investigate T-cell diversity and clonal expansion by sequencing T-cell receptor beta chain rearrangements.
- **AmpliSeq for Illumina Custom RNA Panel** to measure gene expression in 12 to 1200 gene targets of interest in a single assay designed from a menu of > 20,000 human RefSeq genes.

Table 2: AmpliSeq for Illumina targeted RNA panels

	AmpliSeq for Illumina Transcriptome Human Gene Expression Panel	AmpliSeq for Illumina Immune Response Panel	AmpliSeq for Illumina Immune Repertoire Plus, TCR beta Panel	AmpliSeq for Illumina Custom RNA Panel
RNA analyzed	> 95% of human RefSeq genes; 20,802 genes	395 genes associated with immune response	TCRβ chain rearrangements, including CDR1, CDR2, and CDR3	12-1200 custom genes
Method	Targeted RNA-Seq			
Assay time	6 hours	6 hours	5.5-7.5 hours	5.5-7.5 hours
Hands-on time	< 1.5 hours	< 1.5 hours	< 1.5 hours	< 1.5 hours
FFPE support	Yes	Yes	No	Yes
Input quantity	1-100 ng RNA (10 ng recommended)	1-100 ng RNA (10 ng recommended)	10–1000 ng	1 ng
Mechanism of action	Multiplex PCR			
Link	Learn more	Learn more	Learn more	Learn more

Bulk gene regulation solutions

Various sequencing methods are available to investigate gene regulation at the level of DNA/RNA–protein interactions, methylation state, and protein production (Table 3).

DNA/RNA–protein interactions

ChIP-Seq accurately surveys interactions between protein, DNA, and RNA, enabling the interpretation of regulation events central to many biological processes and disease states. Use ChIP-Seq to identify transcription factor binding sites, track histone modifications across the genome, and narrow in on chromatin structure and function.

- **TruSeq™ ChIP Library Prep Kit** provides a simple, cost-effective solution for generating libraries from ChIP-derived DNA for highly multiplexed, cost-effective, and high-quality ChIP-Seq.

Methylation sequencing

Cytosine methylation can significantly modify temporal and spatial gene expression and chromatin remodeling. Genome-wide analysis and targeted NGS approaches can provide researchers with insight into methylation patterns at a single nucleotide level.

- **Whole-genome bisulfite sequencing (WGBS)** is a comprehensive method for interrogating DNA methylation. WGBS relies on the bisulfite conversion of DNA to detect unmethylated cytosines, with read counts used to determine the percentage of methylated cytosines across the genome.
- **TruSeq Methyl Capture EPIC Library Prep Kit** is a targeted methylation sequencing solution for analyzing specific regions of interest in a subset of the genome, producing more manageable data sets and faster workflows compared to WGBS.

Chromatin accessibility

ATAC-Seq is a popular NGS method for determining chromatin accessibility across the genome. By sequencing regions of open chromatin, ATAC-Seq can uncover how chromatin packaging and other factors affect gene expression.[3] ATAC-Seq can be performed on bulk cell populations or on single cells at high resolution.

Table 3: Gene regulation profiling methods

Method	Description	Commercial offering/example method	Link
ChIP-Seq	Surveys interactions between protein, DNA, and RNA by immunoprecipitating DNA–protein complexes and sequencing associated nucleic acids	TruSeq ChIP Library Prep Kit	Learn more
WGBS	Analyzes DNA methylation genome-wide by bisulfite conversion of unmethylated cytosines	Lister R, et al. Human DNA methylomes at base resolution show widesread epigenomic differences. *Nature*. 2009;462(7271):315-322.	Learn more
Targeted methylation sequencing	Interrogates DNA methylation in specific genomic regions of interest by target enrichment after bisulfite conversion and library prep	TruSeq Methyl Capture EPIC Library Prep Kit	Learn more
ATAC-Seq	Assesses chromatin accessibility genome-wide by using a transposase to insert sequencing adapters into regions of open chromatin	Buenrostro J, et al. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29-9	Learn more

Single-cell sequencing solutions

Tissue preparation and cell isolation

Critical steps in single-cell sequencing workflows are the initial tissue preparation and cell isolation prior to library preparation. Early methods for single-cell isolation were low throughput, able to process only dozens to a few thousand cells per experiment. The emerging availability of high-throughput, microfluidic-based methods for cell isolation permits researchers to examine hundreds to tens of thousands of cells per experiment in a cost-effective manner. Researchers can choose from a large ecosystem of tissue preparation, single-cell isolation, and library preparation providers, enabling studies to be tailored to a wide variety of species, tissue/cell types, and methods.

Library preparation

The experimental question will largely determine the cell profiling approach, including library preparation, chosen (Table 4).

Transcriptome profiling

Single-cell RNA sequencing (scRNA-Seq) enables transcriptional profiling of individual cells, which can reveal insights otherwise masked by the mean expression signal of a population of cells sequenced in bulk. Whole transcriptome or targeted gene expression panels are available (Table 4). Advances in immunodetection methods enable combinatorial, multiomic studies combining transcriptome and proteome profiling.

scATAC-Seq

As discussed, ATAC-Seq can elucidate chromatin accessibility genome-wide. Single-cell ATAC-Seq combines compartmentalization and barcoding of single cells with Tn5 (a highly active transposase) tagmentation. The Tn5 transposase tags open chromatin regions with sequencing adapters. The tagged DNA fragments are then purified, amplified, and sequenced.

Protein profiling

Traditionally, protein expression analysis has relied on flow cytometry methods that use fluorophore-conjugated antibodies. Major limitations of this technology include the relatively low numbers of available fluorophores and their spectral overlap. A major advance saw the replacement of the fluorescent label used for protein detection and quantification with oligonucleotide labels, allowing for use of NGS as the readout for protein expression.[4-6] Several methods are available that incorporate DNA-labeled antibodies, including AbSeq, CITE-Seq, and RNA expression and protein sequencing (REAP-Seq).

Learn more by reading the Single-Cell Sequencing Workflow: Critical Steps and Considerations eBook at www.illumina.com/single-cell-rna-sequencing

Table 4: Example amplification techniques for single-cell sequencing

Method	Description	Commercial offering/example method
Gene expression		
Full-length RNA-Seq	Enables amplification of full-length cDNA with Switching Mechanism at 5′ end of RNA Template (SMART) technology	• Takara SMARTer cDNA Synthesis Kits
mRNA end-tag amplification (3′ WTA or 5′ WTA)	Captures mRNA by 3′ polyadenylated (poly(A)) tails to enable sequencing of the coding transcriptome with strand-specific information	• 10x Genomics Chromium Single Cell Gene Expression Solution (3′ WTA) • 10x Genomics Chromium Single Cell Immune Profiling Solution (5′ WTA) • SureCell WTA 3′ Library Prep Kit for the ddSEQ System
Targeted panels	Enables immune receptor (IR), T-cell, breast cancer profiling, and more with various predesigned single-cell targeted RNA sequencing panels	• BD Rhapsody Single-Cell Analysis
Gene regulation		
ATAC-Seq	Assesses chromatin accessibility genomewide by using a transposase to insert sequencing adapters into regions of open chromatin	• 10X Genomics Chromium Single Cell ATAC Solution • Abcam ATAC-Seq protocol • Bio-Rad SureCell ATAC-Seq Library Prep Kit
Protein profiling		
AbSeq	Enables protein profiling by NGS with DNA-tagged antibodies	• BD AbSeq antibody-oligonucleotide conjugates
CITE-Seq	Integrates cellular protein and transcriptome measurements into a single assay using oligonucleotide-labeled antibodies	• Stoeckius M, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14:865-868. • cite-seq.com
REAP-Seq	Quantifies gene and protein expression levels with DNA-antibody conjugates	• Peterson VM, et al. Mulitplexed quantification of proteins and transcripts in single cells. *Nat Biotech*. 2017;35:936-939.

Choosing NGS instrumentation

Adopting an NGS system is an important decision for any lab. Illumina offers innovative NGS platforms that deliver exceptional data quality and accuracy, at a massive scale for both bulk and single-cell sequencing experiments. With a broad range of options available, depending on budget, the nature of research, and specific experimental objectives, there is a system that will meet virtually every lab's needs (Table 5).

Table 5: Illumina sequencing systems at a glance

						
System	**iSeq 100 System**	**MiniSeq System**	**MiSeq System**	**NextSeq 550 System**	**NextSeq 1000 and NextSeq 2000 Systems**	**NovaSeq 6000 System**
Most important to me	Affordability and efficiency	Simplicity and instrument affordability	Speed, accuracy, and simplicity	Flexible desktop sequencing system for exome, transcriptome, and whole-genome sequencing	High-throughput sequencing power and flexibility to scale based on project or workflow needs	Highthroughput, low-cost sequencing for production-scale genomics
Instrument control software	Local Run Manager	Local Run Manager	Local Run Manager	Local Run Manager	Local Run Manager	Illumina Experiment Manager
Onboard informatics	Yes	Yes	Yes	No	Yes	No
Benchtop system	Yes	Yes	Yes	Yes	Yes	No
Production-scale capabilities	No	No	No	Yes	Yes	Yes
Flow cell options	Standard	Mid-output, High-output	Standard v2, Micro v2, Nano v2, Standard v3	Mid-output, High-output	P2, P3	SP, S1, S2, S4
Flow cells processed/run	1	1	1	1	1	1 or 2

Benchtop low-throughput sequencing systems

Benchtop sequencing systems are the most affordable NGS instrumentation option in terms of initial purchase price. They are ideal for small- to medium-scale gene expression and regulation studies. These sequencing systems provide fast workflows for targeted samples. A common application is targeted gene expression, which allows sequencing of a subset of mRNA transcripts or specific genomic regions of interest efficiently and cost-effectively. Some larger labs choose to have benchtop systems in addition to their high-throughput equipment for small-scale follow-up studies, for use in library quality control (QC), and more.

iSeq™ 100 System

With the lowest price, smallest footprint, and fastest run time of any Illumina instrument, the iSeq 100 System offers an affordable option for researchers to include NGS in their research. Sequence targeted genes, RNA transcripts, and more at the push of a button.[7]

Learn more at www.illumina.com/iseq

iSeq 100 System flow cell options and specifications	
Flow cell type	i1
Output/run	144 Mb-1.2 Gb
Reads/run	4M
Max read length	2 × 150 bp
Cycles	300

MiniSeq™ System

The MiniSeq System offers a simple, affordable solution for a broad range of targeted DNA and RNA sequencing applications for examining single genes or entire pathways. An intuitive user interface, load-and-go operation, and onboard data analysis make it easy to learn and easy to use.[8]

Learn more at www.illumina.com/miniseq

MiniSeq System flow cell options and specifications				
Flow cell type	Mid-output	High-output		
Output/run	2.1-2.4 Gb	1.7-1.9 Gb	3.3-3.8 Gb	6.6-7.5 Gb
Reads/run	8M	25M	25M	25M
Max read length	2 × 150 bp	1 × 75 bp	2 × 75 bp	2 × 150 bp
Cycles	300	75	150	300

MiSeq™ System

The MiSeq System combines speed, high-quality data, and the longest read lengths from Illumina. The MiSeq System is ideal for a diverse range of targeted RNA gene expression and regulation studies and is suitable for moderate sample batches.[9]

Learn more at www.illumina.com/miseq

MiSeq System flow cell options and specifications								
Flow cell type	Standard v2			Micro v2	Nano v2		Standard v3	
Output/run	0.75-0.85 Gb	4.5-5.1 Gb	7.5-8.5 Gb	1.2 Gb	0.3 Gb	0.5 Gb	3.3-3.8 Gb	13.2-15 Gb
Reads/run	15M	15M	15M	4M	1M	1M	25M	25M
Max read length	2 × 25 bp	2 × 150 bp	2 × 250 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp	2 × 75 bp	2 × 300 bp
Cycles	50	300	500	300	300	500	150	600

Benchtop high-throughput sequencing systems

Benchtop high-throughput NGS instruments have a mid-range price point while offering accessibility and convenience. These higher throughput systems maintain a small-footprint benchtop format. They are commonly used for exome, mRNA, and single-cell sequencing studies.

NextSeq™ 550 System

The NextSeq 550 System delivers the power of high-throughput sequencing with the speed, simplicity, and affordability of a benchtop NGS system. It supports mid- to high-throughput sequencing applications and is ideal for smaller scale single-cell sequencing studies. The NextSeq 550 System is efficient and flexible to handle a range of different types of projects, delivering on-demand transcriptomics at both bulk and single-cell levels with a rapid, one-day turnaround.[10]

Learn more at www.illumina.com/nextseq550

NextSeq 550 System flow cell options and specifications					
Flow cell type	Mid-output v2.5		High-output v2.5		
Output/run	16-19 Gb	32-39 Gb	25-30 Gb	50-60 Gb	100-120 Gb
Reads/run	130M	130M	400M	400M	400M
Max read length	2 × 75 bp	2 × 150 bp	1 × 75 bp	2 × 75 bp	2 × 150 bp
Cycles	150	300	75	150	300

NextSeq 1000 and NextSeq 2000 Systems

The NextSeq 1000 and NextSeq 2000 Sequencing Systems use the latest advances to miniaturize the volume of the sequencing reaction while increasing output and reducing the cost per run. Users can obtain the throughput, data quality, and cost needed to expand the size and scope of their studies on a benchtop sequencing system. With improved technology, advanced chemistry, simplified workflows, and onboard secondary analysis, researchers have unprecedented flexibility to investigate and discover. The NextSeq 2000 System provides higher throughput options to meet the needs of new and emerging applications while achieving better run economics for current applications. The NextSeq 1000 System has lower throughput relative to the NextSeq 2000 System and is available at a lower system price. To ensure flexible future scalability, customers who purchase a NextSeq 1000 System can easily upgrade to the NextSeq 2000 System [11]

Learn more at www.illumina.com/nextseq2000

NextSeq 2000 System flow cell options and specifications						
Flow cell type	NextSeq 1000/2000 P2 Reagents			NextSeq 2000 P3 Reagents		
Output/run	40 Gb	80 Gb	120 Gb	100 Gb	200 Gb	300 Gb
Reads/run	400M	400M	400M	1B	1B	1B
Max read length	2 × 50 bp	2 × 100 bp	2 × 150 bp	2 × 50 bp	2 × 100 bp	2 × 150 bp
Cycles	100	200	300	100	200	300

High-throughput/high-volume sequencing system

For busy labs, a high-throughput/high-volume sequencing system delivers the lowest price per sample. They can complete projects in the least number of runs for the highest operational efficiency. Single-cell and bulk applications include whole-transcriptome sequencing and sequencing regarding epigenetic regulation, as well as exome and mRNA sequencing.

NovaSeq™ 6000 System

The NovaSeq 6000 System represents the most powerful, scalable, and reliable high-throughput Illumina sequencing platform, producing outstanding data quality. With the highest throughput of any production platform, the NovaSeq 6000 System enables users to power studies rapidly with more samples and higher depth of coverage, making it ideal for extensive screening studies, such as pharmaceutical screens, single-cell atlas studies, and other large-scale experiments.[12]

Learn more at www.illumina.com/novaseq

The NovaSeq 6000 v1.5 Reagent Kit

Illumina offers the NovaSeq 6000 v1.5 Reagent Kit. This kit provides improved economics, more flexible sequencing workflows, and extended reagent shelf-life. The NovaSeq 6000 v1.5 Reagent Kit allows labs to elevate their NGS capabilities, while maintaining the same high data quality previously obtained with v1.0 reagents.

Learn more by reading the *Enhanced sequencing capabilities with the NovaSeq 6000 v1.5 Reagent Kit* technical note

NovaSeq 6000 System flow cell options and specifications												
Flow cell type	SP (v1.5 reagents)			S1 (v1.5 reagents)			S2 (v1.5 reagents)			S4 (v1.5 reagents)		
Output/run	80 Gb	250 Gb	400 Gb	167 Gb	333 Gb	500 Gb	417 Gb	833 Gb	1250 Gb	350 Gb	2000 Gb	3000 Gb
Reads/run	800M	800M	800M	1600M	1600M	1600M	4100M	4100M	4100M	8000-10,000M		
Cycles	100	300	500	100	200	300	100	200	300	35	200	300
Output per flow cell												
1 × 35 bp	N/A			N/A			N/A			280-350 Gb		
2 × 50 bp	65-80 Gb			134-167 Gb			333-417 Gb			N/A		
2 × 100 bp	134-167 Gb			266-333 Gb			667-833 Gb			1600-2000 Gb		
2 × 150 bp	200-250 Gb			400-500 Gb			1000-1250 Gb			2400-3000 Gb		
2 × 250 bp	325-400 Gb			N/A			N/A			N/A		

N/A = not applicable

Technical considerations for sequencing

Before choosing an NGS system, there are several factors to consider, including read depth, single- or paired-end reads, quality metrics, and instrument control software.

Sequencing coverage (read) depth

Sequencing coverage for traditional or bulk samples describes the average number of reads that align to, or "cover," known reference bases. The sequencing coverage level (or "read depth") often determines the degree of confidence of variant discovery at particular base positions.

Sequencing coverage requirements vary by application. At higher levels of coverage, each base is covered by a greater number of aligned sequence reads, so base calls can be made with a higher degree of confidence (Figure 1).[13]

For various single-cell sequencing applications, read depth is discussed not in the number of reads per base, but in the number of reads per cell. The required sequencing depth for a single-cell sequencing run will depend on several factors, including sample type, the number of cells to be analyzed, experimental objectives, and more. Ultimately, the required sequencing depth will largely depend on sample type and experimental objective and will need to be optimized for each study.

Learn more at www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html

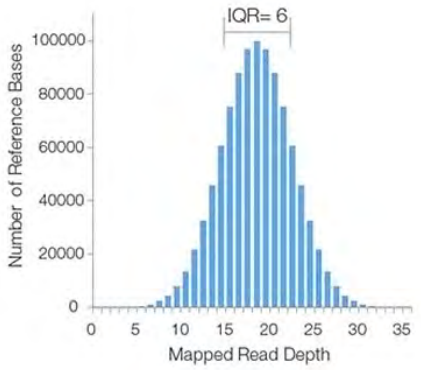


Figure 1: Sequencing coverage histogram—Example plot in the form of a Poisson-like distribution with a small standard deviation, valid under the assumption that reads are randomly distributed across the genome and that the ability to detect true overlaps between reads is constant within a sequencing run.

Sequencing read options

Sequencing can involve single-read or paired-end reads. Single-read sequencing of DNA proceeds from only one end, and is the simplest way to use Illumina sequencing. It delivers large volumes of high-quality data, faster and cheaper than paired-end sequencing.[14] Single-read runs can be a good choice for certain methods such as small RNA sequencing. In contrast, paired-end sequencing involves sequencing both ends of DNA fragments in a library and aligning the forward and reverse reads as read pairs. This results in better alignment of reads, especially across repetitive, difficult-to-sequence regions (Figure 2). All Illumina NGS systems are capable of paired-end sequencing.

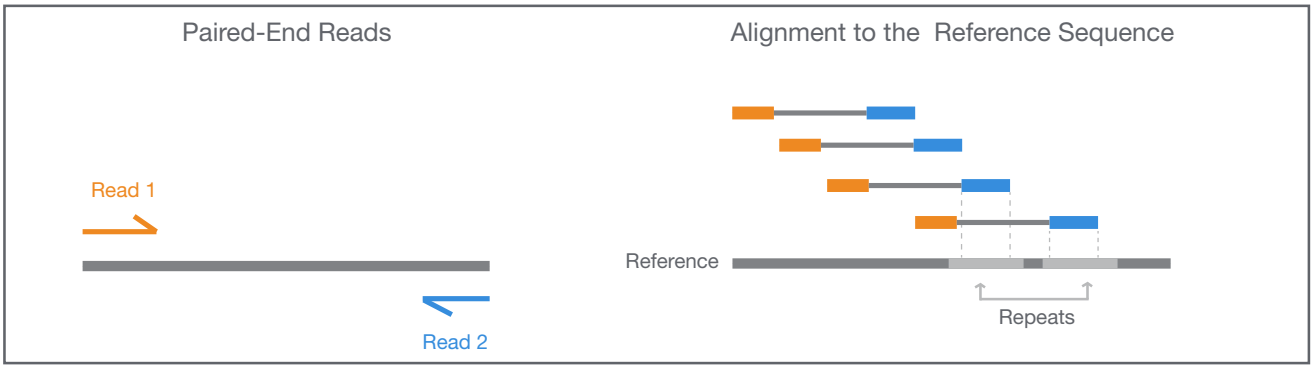


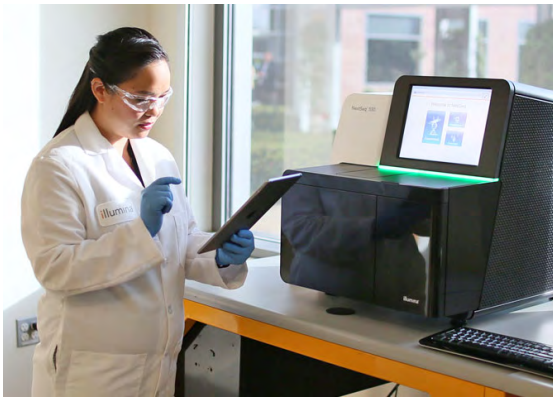
Figure 2: Paired-end sequencing and alignment—Paired-end sequencing enables sequencing of both ends of DNA fragments. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely than single-read sequencing.

In addition to producing twice the number of reads for the same time and effort in library preparation, sequences aligned as read pairs enable detection of insertion-deletion (indel) variants, which is not possible with single-read data.[15] Paired-end RNA sequencing enables discovery applications such as detecting gene fusions, novel transcripts, and novel splice isoforms.[16]

NGS quality metrics

Sequencing quality metrics can provide important information about the accuracy of each step in the NGS workflow. Base calling accuracy, measured by the Phred quality score (Q-score), is a common metric used to assess the accuracy of a sequencing platform. It indicates the probability that a given base is called incorrectly by the sequencing system.

A Q30 score is the percentage of bases with a quality score of 30 or higher, which means that the probability for an error in base calling is 1 in 1000 and the base calling accuracy is 99.9%. Illumina sequencing by synthesis (SBS) chemistry delivers an exceptionally high percentage of error-free reads, with most bases having quality scores above Q30 (% ≥ Q30) across all Illumina sequencing systems.



Instrument control software

Many Illumina sequencing systems come preinstalled with Local Run Manager, an integrated on-instrument solution for creating a run, monitoring status, analyzing sequencing data, and viewing results. Local Run Manager minimizes setup and the risk of user error. The software enables onboard review of real-time data and performance metrics. Local Run Manager can be accessed through the user-friendly on-instrument interfaces or via a web browser.

The NovaSeq 6000 System uses Illumina Experiment Manager to design experiments before an Illumina sequencing run. The software guides users through the creation and setup of a sample sheet. The built-in validation checks help minimize errors by detecting and warning of suboptimal index combinations.

BaseSpace™ Clarity LIMS is a laboratory information management system (LIMS) that helps any lab track samples, manage workflows, and streamline operations for optimized, efficient sequencing. BaseSpace Clarity LIMS integrates easily with all Illumina instruments.

Summary

Different gene expression and regulation experiment types, whether bulk or single-cell, have specific library preparation and sequencing requirements. Illumina offers robust library prep solutions for various bulk RNA-Seq applications. Illumina supports third-party scRNA-Seq library prep solutions with compatible sequencing systems that offer high data accuracy with flexible throughput. Together, Illumina delivers proven NGS solutions for gene expression and regulation studies at any scale. After high-quality, reliable RNA-Seq data are obtained, researchers can proceed with data visualization, analysis, and interpretation.

Chapter 2: The bioinformatics workflow for NGS-based gene expression and regulation studies

NGS-based gene expression and regulation experiments generate large volumes of raw data. Analyzing this data can be both intimidating and challenging. Fortunately, the data storage and analytical tools available today take much of the manual work out of the data analysis process. User-friendly Illumina bioinformatics tools smooth the analysis process, making it easier for researchers to gather meaningful biological information quickly.

The NGS data analysis workflow consists of three main steps: primary, secondary, and tertiary analysis (Figure 3). Primary data analysis consists of the digitization of genetic information into "reads" of nucleotide sequences and the conversion of raw sequencing files into a more versatile format. During secondary analysis, sequencing reads are demultiplexed (if applicable) and aligned to an annotated reference genome (or assembled *de novo*, if a reference genome is not available) to enable genetic characterization, which can include measuring quantities of genes and/or transcripts within a sample. Tertiary analysis can vary depending on the experimental design goals of the researcher, but generally involves data interpretation in a broader biological context. It can include variant annotation, filtering, prioritization, data visualization, and reporting.

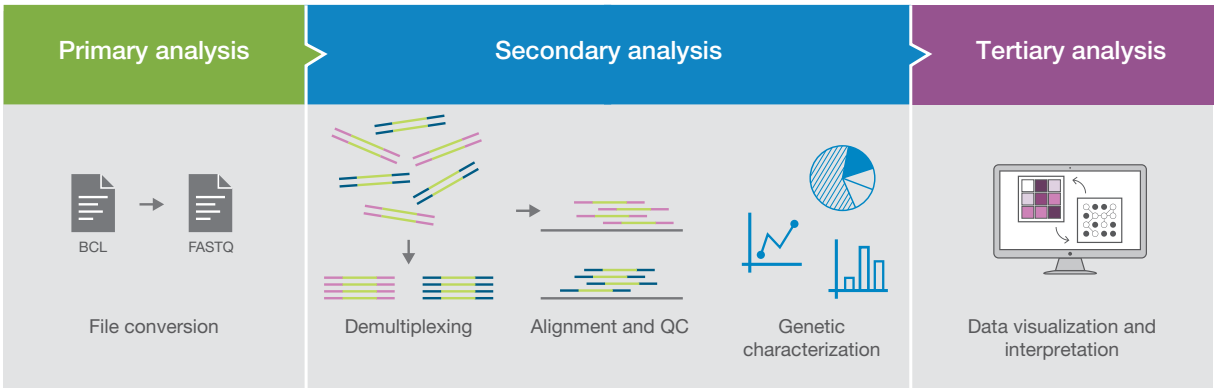


Figure 3: The NGS data analysis workflow—NGS data analysis includes three main steps. In primary analysis, raw base call (BCL) files are converted to text-based sequence (FASTQ) files. During secondary analysis, reads are demultiplexed and aligned to a reference genome. The resulting sequence undergoes basic genetic characterization. Tertiary analysis involves advanced data visualization and biological interpretation.

Primary analysis—file conversion

Primary analysis begins on all Illumina sequencing systems, with on-instrument Real-Time Analysis (RTA) software operating during cycles of SBS chemistry and imaging. RTA software provides base calls and associated quality scores representing the primary structure of DNA or RNA strands. Raw output of a sequencing run is stored as individual base call (BCL) files. When sequencing completes, BCL files must be converted into FASTQ format for use with downstream analysis tools. FASTQ is a text-based sequence file format that stores both raw sequence data and quality scores. The inclusion of quality scores makes FASTQ files an important part of sequencing QC, as this information will be used to filter and discard underperforming reads from downstream analysis. FASTQ files have become the standard format for storing NGS data from Illumina sequencing systems and can be used as input for a wide variety of secondary data analysis solutions.

Secondary analysis—demultiplexing, alignment and QC, and genetic characterization

After the sequencing run is complete and raw read files are converted to FASTQ files, researchers can proceed with secondary analysis. NGS data are demultiplexed and aligned to a reference genome. Various data QC metrics can be used to assess the quality of the data before downstream analysis.

Demultiplexing

A key component of the increased capacity of Illumina sequencing systems is multiplexing, which adds unique sequences, called indexes, to each DNA fragment during library preparation. This allows large numbers of libraries to be pooled and sequenced simultaneously during a single sequencing run. Demultiplexing separates sequencing reads from pooled libraries into individual libraries based on the index sequences. For single-cell sequencing, demultiplexing involves separating reads from pooled samples into individual cells based on cell barcodes or unique molecular identifiers (UMIs) added during cell isolation.

Alignment

Alignment maps FASTQ files to a reference genome. In the absence of a relevant reference genome, reads are assembled into longer contiguous segments called "contigs." Various software applications are available to perform sequence alignment, including the Burrows-Wheeler Alignment (BWA)[17] algorithm, used in the BWA Aligner BaseSpace App, and Spliced Transcripts Alignment to a Reference (STAR)[18] algorithm, included in the RNA-Seq Alignment BaseSpace App (Table 6).

Table 6: BaseSpace Apps sequence alignment

BaseSpace app	Description	Link
 BWA Aligner	The BWA Aligner App aligns samples (consisting of FASTQ files) using the BWA-MEM aligner to a reference genome, including a custom reference genome created from imported FASTA files.	Learn more
 RNA-Seq Alignment	The RNA-Seq Alignment App performs the following: read mapping using the STAR aligner, quantification of reference genes and transcripts using Salmon, variant calling (SNVs and small indels) using the Strelka Variant caller, fusion calling with Manta, and QC metrics from Picard and other sources.	Learn more

QC metrics

QC is a critical component of any NGS experiment, as it increases confidence in the accuracy and reproducibility of the results. QC metrics can include the quality of the input RNA, the raw read data generated by the sequencing system, and mapping and alignment of the sequencing reads.

Raw read data QC

Common QC metrics used to evaluate raw sequencing read data include GC content, sequencing Q-scores, representation of sequences, and k-mer. GC content, calculated as the percentage of guanine (G) + cytosine (C) in a reference sequence, can be used to approximate the expected GC content of the generated sequence. Significant deviation of GC content in the sequence data from the reference can indicate contamination in the sample. For RNA-Seq, the GC content varies by RNA type (Table 7). Expected GC content can be used as a QC metric, depending on the specific RNA-Seq method, eg, 39.7–48.9% for total RNA-Seq.[21]

Table 7: Expected GC content for human RNA types

RNA type	Expected GC content
Coding RNA	48.9%
Long noncoding RNA	39.7%
rRNA	50.2%
miRNA	51.5%
Transfer RNA (tRNA)	55.7%
Other small RNAs	46.7%

As discussed in the previous chapter, Q-scores can be used to assess the accuracy of base calling by the instrument, with a Q30 score indicating accuracy of 99.9%. All Illumina sequencing systems perform at ≥ 75% of bases higher than Q30, averaged across an entire sequencing run. For RNA-Seq, additional metrics are valuable

for evaluating raw read data, regarding gene expression levels. High-quality sequencing libraries have a diverse composition of RNA sequences; therefore, overrepresentation of particular sequences may indicate contamination from sequencing adapters or other sources. Sequence representation analysis can overlook shorter sequences (< 10 nucleotides). K-mer analysis examines all possible nucleotide combinations of a length, k, to check for short, duplicated sequences.

Alignment QC

For RNA-Seq, a key aspect of read alignment is capture efficiency, which refers to the percentage of total sequenced reads that map to the target region of interest. For NGS methods, capture efficiency is not 100%, and can range from 50-80% for methods such as exome sequencing and RNA-Seq.[22] Generally, higher quality samples result in higher capture efficiencies. Low capture efficiencies observed in RNA-Seq data can indicate low sample quality, contamination by other sample RNA, or, in the case of total RNA-Seq, inefficient removal of rRNAs.

RNA quality and integrity

RNA quality and integrity are crucial for successful, high-quality RNA-Seq results. RNA integrity can be measured by evaluating the ratio of two ribosomal RNAs (rRNAs), 28s:18s, where a higher ratio generally indicates higher quality RNA with better integrity. Analysis of RNA samples with a Bioanalyzer system (Agilent) or Fragment Analyzer system (Agilent) can produce an RNA Integrity Number (RIN) or RNA Quality Number (RQN), respectively, two equivalent metrics to evaluate RNA integrity.[19,20]

To learn more about assessing RNA integrity, read the *Scalable Nucleic Acid Quality Assessments for Illumina NGS Preparation* application note.

Genetic characterization

NGS data can be instantly and securely transferred, stored, and analyzed in BaseSpace Sequence Hub, the Illumina genomics cloud-computing platform. In addition, the Illumina DRAGEN™ Bio-IT Platform provides accurate, ultra-rapid secondary analysis of NGS data, in BaseSpace Sequence Hub or on-premise.

BaseSpace Sequence Hub

BaseSpace Sequence Hub can be accessed from any internet-connected computer or mobile device and offers a security-first environment that enables any researcher to set up runs and monitor instrument run quality. By simplifying storage and analysis of sequence data, BaseSpace Sequence Hub can reduce the need for capital expenditure on in-house computing infrastructure.[23] For more sophisticated users, BaseSpace Sequence Hub can integrate with in-house informatics systems via extensible application programming interfaces (APIs).

BaseSpace Sequence Hub offers various push-button analysis solutions and version-controlled, ready-to-use workflows. It also supports simple, secure, and efficient pipeline configuration for personalized analysis workflows. The BaseSpace Apps store offers a wide variety of tools that are developed or optimized by Illumina, or from a growing network of third-party app providers (Figure 4).

Learn more at www.illumina.com/basespace

DRAGEN Bio-IT Platform

The DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Platform uses highly reconfigurable field-programmable gate array technology (FPGA) to provide hardware-accelerated implementations of genomic analysis algorithms, such as BCL conversion, mapping, alignment, sorting, duplicate marking, and haplotype variant calling. Fundamental features of the DRAGEN Platform address common challenges in genomic analysis, such as lengthy compute times and massive volumes of data.[24] The reprogrammable nature of the DRAGEN platform enables Illumina to develop custom algorithms and allows for improvements to accommodate future applications.

Learn more at www.illumina.com/DRAGEN

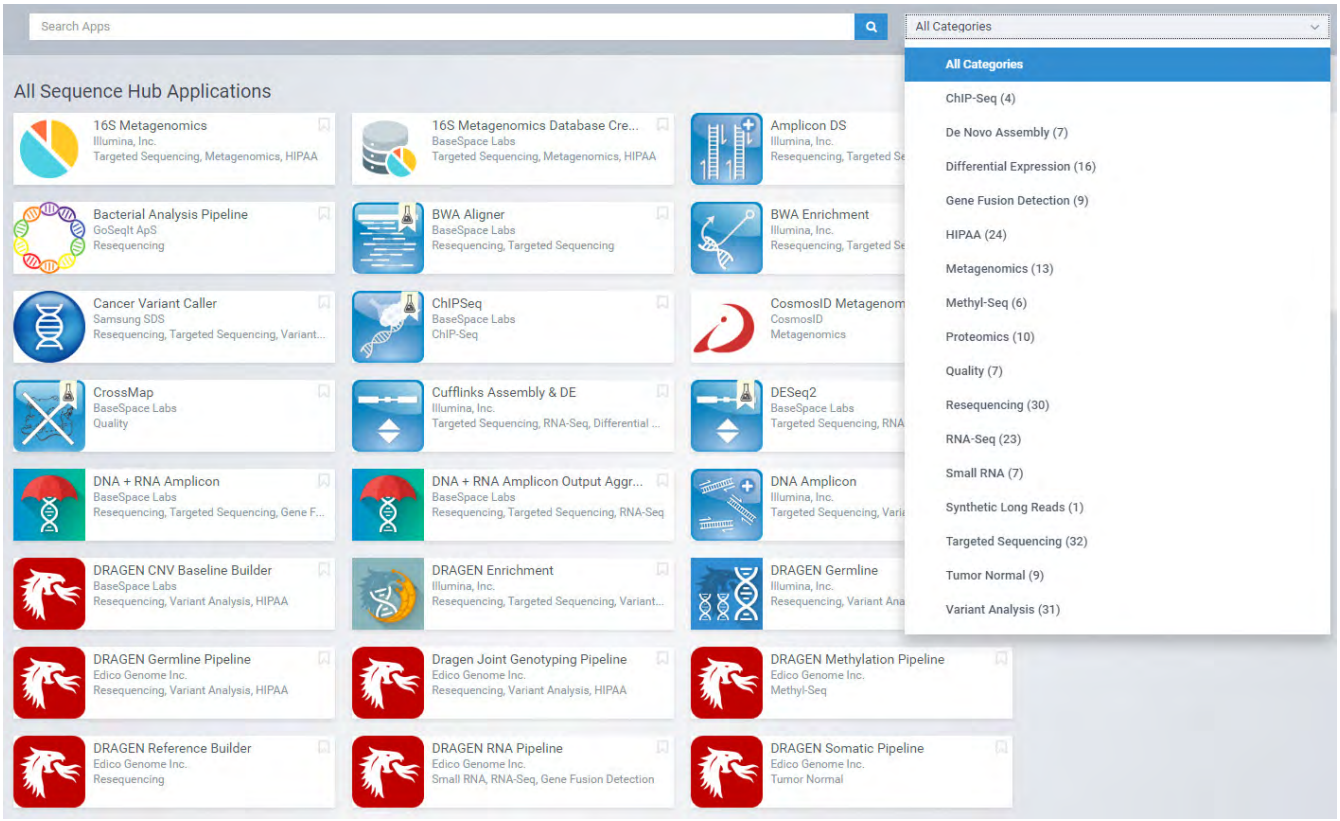


Figure 4: Analytical tools on demand—Researchers can browse and explore a growing list of apps from the bioinformatics community in the BaseSpace Apps store, including DRAGEN pipelines, and launch selected apps with a single click, directly from the data set.

Tertiary analysis—data visualization and interpretation

The goal of tertiary analysis is to add biological context to the results generated from secondary analysis. Depending on the biological question and sequencing method, tertiary analysis can diverge into a spectrum of various study-specific downstream investigations. For example, differential gene expression analysis will start with gene expression quantification for each individual sample, followed by a comparison between groups to identify genes or transcripts exhibiting statistically different levels of abundance. This data can also be combined with phenotype information, biological knowledge from functional genomics, or other data sources to better understand the broader impact of the secondary analysis results. Secondary analysis results may also be combined with additional experimental data using other genomic data types to provide a multiomic view towards biology, including epigenetic and proteomic approaches. There are many software programs and tools that aid in data interpretation by providing advanced, interactive data visualization and exploration (Figure 5).

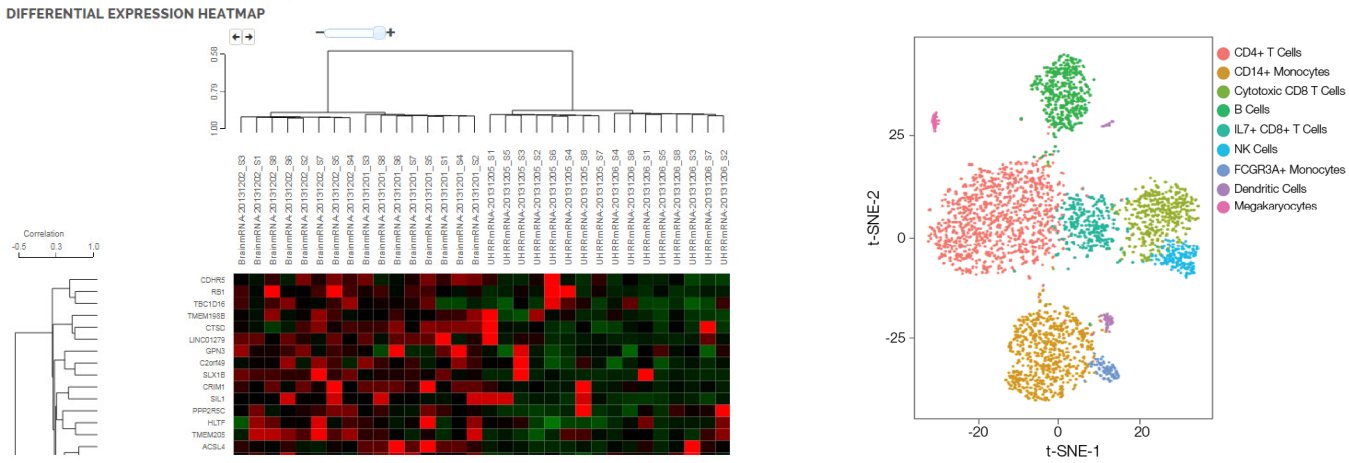


Figure 5: Advanced data visualization—Various software tools enable visualization and exploration of NGS data, specific to study type and NGS method used. Examples include heat maps to visualize differential gene expression (left) and cell cluster identification to explore single-cell sequencing data (right). The example heat map was generated with the RNA-Seq Differential Expression BaseSpace App. The cell cluster plot was generated with Seurat software.

Applying biological interpretation methods allows researchers to derive insights into fundamental biological processes and the causes of genetic disease. From sequence data, bioinformatics software and biological data mining approaches convert data into knowledge. Illumina offers BaseSpace Correlation Engine, a bioinformatics solution that connects new experimental data sets with a large, curated repository of open-access and controlled-access public NGS and microarray data (Figure 6). This interactive data analysis environment helps researchers discover new associations to diseases, tissues, and literature using a data-driven approach to correlate new data to existing studies.[25]

Learn more about BaseSpace Correlation Engine at www.illumina.com/products/by-type/informatics-products/basespace-correlation-engine.html

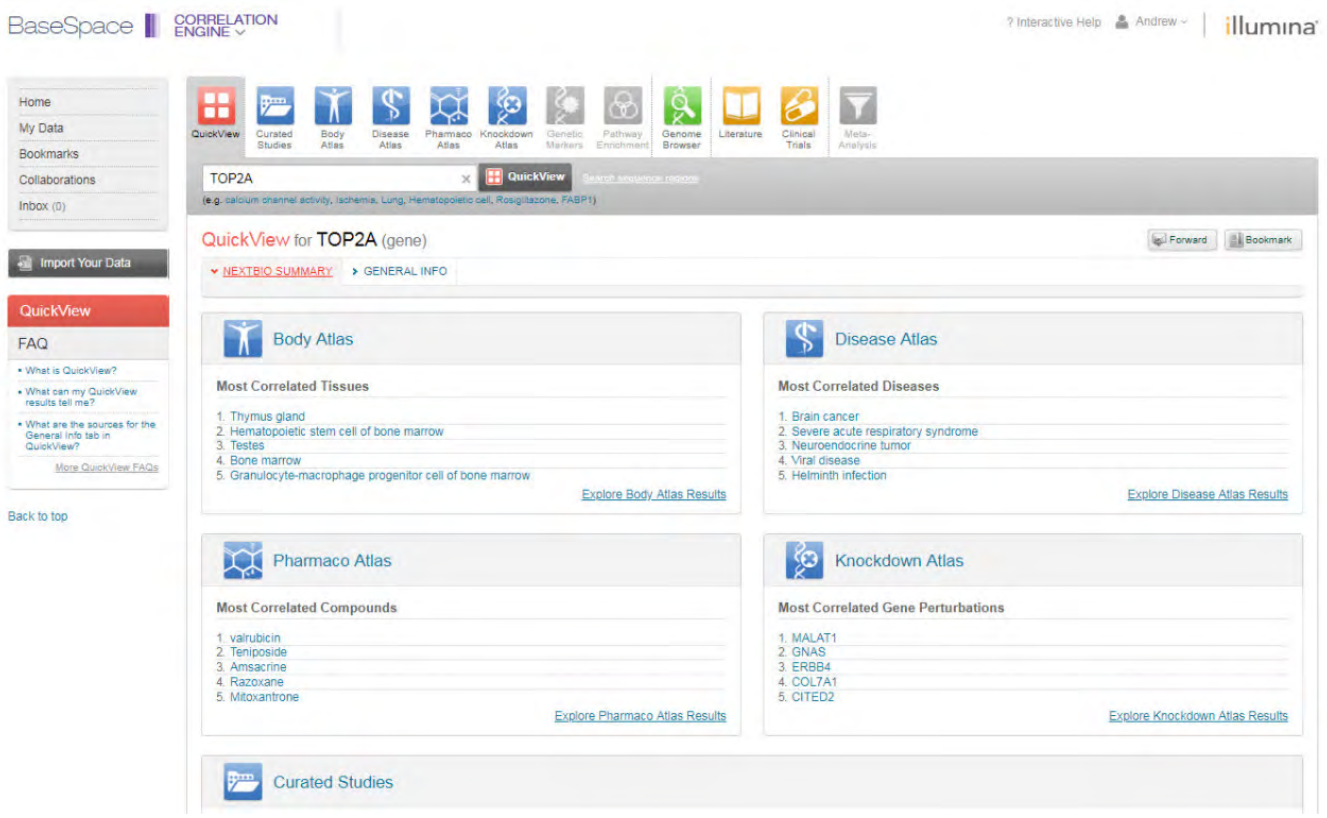


Figure 6: BaseSpace Correlation Engine—The software interface enables quick idenitfication of novel correlations and associations for a given query, revealing data-driven connections between genes, diseases, compounds, tissues, pathways, and literature.

Summary

Researchers performing NGS-powered gene expression and regulation experiments have a robust and growing landscape of bioinformatics analysis tools available for a wide variety of sample types and biological questions. From basic gene quantification and differential expression analysis, to *de novo* transcriptome assembly methods, NGS is continually evolving to enable a better understanding of the biology underlying gene expression, regulation, and transcription into protein. The NGS analysis workflow proceeds from file conversion (primary) to alignment and genetic characterization (secondary) to advanced data visualization and biological interpretation (tertiary). Depending on the experimental objective and sequencing method, there is an extensive array of bioinformatics tools and pipelines available to facilitate secondary and tertiary analysis.

Chapter 3: Bioinformatics pipelines for NGS-based gene expression and regulation studies

NGS-based bulk gene expression and regulation studies produce large, complex data sets that require fast, scalable, user-friendly analysis software to derive meaningful insights into the underlying biology. Currently available bioinformatics tools and pipelines can be used by all researchers, even those without prior bioinformatics experience, to analyze data. This chapter covers the NGS bioinformatics workflow for secondary and tertiary analysis of bulk RNA-Seq and ATAC-Seq.

Bioinformatics solutions for bulk gene expression studies

Genetic characterization of RNA-Seq gene expression studies quantifies the abundance of transcripts at each gene position using an annotated reference genome to determine gene positions and transcript identities. For differential expression analysis, output files will contain a row for each gene or transcript position, a column for each sample or group, and a p-value or q-value reporting the statistical significance of the difference in expression. If the analysis included alignment, a BAM file is produced to report where each read aligned to the reference genome. For samples without a relevant reference genome, secondary analysis may consist of multiple steps, including genome assembly to provide a scaffold genome upon which transcripts can then be quantified. In addition to quantifying transcript abundance, variants can also be identified and quantified. SNVs and indels will be reported in a VCF or genome VCF file. Splice variants may also be included as a separate VCF file.

Several analysis tools, varying in their approaches with relative advantages and disadvantages, are available for secondary analysis for gene expression (Table 8). For example, many analysis pipelines begin with alignment (eg, TopHat, STAR, Strelka), but alignment-free approaches (eg, Sailfish, Salmon) that are more computationally efficient, but may exhibit lower accuracy with lower-abundance transcripts are also available.[26] Approaches may also differ in how the tools map reads that ambiguously align to more than one place in a genome or how they quantify abundance when transcript isoforms are present.

TopHat2 and Cufflinks

TopHat2 and Cufflinks are software tools for gene discovery and comprehensive expression analysis of NGS data, including data from single-cell experiments that use full-length methods.[27] TopHat2 performs sequencing read alignment and splice site discovery. Cufflinks assembles read alignments from TopHat2 into transcripts. Together, these apps enable biologists to identify new genes and novel splice variants within known genes and compare gene and transcript expression under multiple conditions. Simple-to-follow prompts guide users through the entire process, starting from selecting the files generated by the sequencing system to filtering and analyzing the NGS data.

STAR

STAR aligner is a fast RNA-Seq read mapper with support for splice junction and fusion read detection. During alignment, different parts of a read are mapped to different genomic positions corresponding to splicing or RNA-fusions. STAR includes known splice junctions from annotated gene models, allowing for sensitive detection of spliced reads.

Salmon

Salmon is a software tool for quantifying gene expression. Salmon combines new algorithms with bias-aware models to accurately measure transcript abundance across the transcriptome.

Sailfish

Sailfish is a software program for quantifying RNA isoforms that have been previously annotated in RNA-Seq data. Sailfish omits read mapping, resulting in significant time savings for reanalyzing sequencing data.

Strelka

The Strelka Somatic Variant Caller analyzes aligned sequencing reads to identify SNVs and indels. It is ideal for variant detection in tumor–normal analysis. Strelka uses an algorithm that enables highly sensitive variant calling with low-purity tumor samples.

Table 8: RNA-Seq analysis tools

Software program	Description	Link
Bowtie2	Aligns short reads	Learn more
TopHat2	Aligns RNA-Seq reads, discovers splice sites	Learn more
TopHat-Fusion	Discovers gene fusions	Learn more
Cufflinks	Assembles transcripts	Learn more
Cuffcompare	Compares assemblies to a reference	
Cuffmerge	Combines multiple assemblies	
Cuffdiff	Performs differential expression analysis	
Strelka	Calls small variants	Learn more
STAR	Maps RNA-Seq reads, detects splice-junctions and gene fusions	Learn more
Salmon	Quantifies transcript expression	Learn more
Sailfish	Quantifies RNA isoforms	Learn more

BaseSpace Apps for RNA-Seq analysis

Illumina packages the aforementioned RNA-Seq tools and additional tools into applications (apps) available on BaseSpace Sequence Hub (Table 9). For example, the RNA-Seq Alignment and RNA-Seq Differential Expression apps provide functionality to align reads, quantify gene and transcript abundance, call variants for SNVs and small indels, call gene fusion candidates, and provide QC metrics for users. Additionally, DRAGEN pipelines are available in BaseSpace Sequence Hub and on the local DRAGEN Server for rapid, highly accurate secondary analysis of RNA transcripts.

Analysis of targeted RNA-Seq using AmpliSeq for Illumina

Sequencing data generated using AmpliSeq for Illumina panels can be analyzed with the RNA Amplicon BaseSpace App (Table 9). RNA Amplicon enables streamlined gene expression analysis using DESeq2 and alignment performed using the BWA. This app supports custom amplicon panel analysis via import of custom manifests. The same analysis workflows are available for on-premise use through Local Run Manager software on select sequencing systems. Analysis results can be stored or easily shared with other investigators. Further analysis can be performed on any variant calls using BaseSpace Variant Interpreter, an interpretation and reporting platform designed to decrease the time and effort required to extract biological insight from genomic data while maximizing operational efficiency (Figure 7).

Learn more about BaseSpace Variant Interpreter at www.illumina.com/products/by-type/informatics-products/basespace-variant-interpreter.html

Table 9: BaseSpace Apps for RNA-Seq analysis

BaseSpace App	Description	Link
Gene expression		
 Cufflinks Assembly & DE	Cufflinks Assembly & DE quickly assesses novel transcript isoforms and gene expression levels, based on analysis results from the RNA-Seq Alignment App. The app uses Cuffmerge, Cuffquant, Cuffnorm, and Cufflinks tools to perform novel transcript merging and differential expression analysis.	Learn more
 RNA-Seq Alignment	RNA-Seq Alignment performs read mapping using STAR, quantification of reference genes and transcripts using Salmon, variant calling using Strelka, and fusion calling with Manta.	Learn more
 RNA-Seq Differential Expression	RNA-Seq Differential Expression performs differential expression analysis of reference genes and transcriptome mining using STAR or TopHat for alignment, gene and transcript counts, annotation, variant calling, fusion detection, and novel transcript assembly.	Learn more
 DESeq2	DESeq2 performs differential expression analysis of reference genes on aligned samples to produce gene counts, gene FPKMs, principal component analysis, and control vs comparison results.	Learn more
 RNA Express	RNA Express combines the capabilities of the STAR aligner and DESeq2 analysis tools in one simple workflow to provides the most commonly used set of RNA analysis features in a convenient and rapid analysis package.	Learn more
 DRAGEN RNA Pipeline	DRAGEN RNA Pipeline performs secondary analysis of RNA transcripts. It offers multiple operating modes, including reference-only alignment and annotation-assisted alignment with gene fusion detection. The gene fusion module leverages the DRAGEN RNA Spliced Aligner to perform split-read analysis on supplementary (chimeric) alignments to detect potential breakpoints, while adding minimal processing time to the overall pipeline.	Learn more
 DRAGEN Differential Expression	DRAGEN Differential Expression performs secondary analysis of RNA transcripts. It runs the DESeq2 algorithm on Salmon quantification files to output genes and transcripts that are differentially expressed between two sample groups.	Learn more
 RNA Amplicon	RNA Amplicon enables streamlined gene expression analysis of NGS amplicon panels, including alignment and differential expression analysis. This app supports custom amplicon panel analysis via import of custom manifests.	Learn more
Gene regulation		
 ChIPSeq	ChIPSeq identifies enriched regions pulled down by ChIP and discovers motifs within these regions. Raw outputs of ChIPSeq are made available for download, and are also presented in an interactive table.	Learn more
 MethylSeq	MethylSeq rapidly analyzes whole-genome and targeted bisulfite DNA sequence data. It performs alignment, methyl calling, and calculates alignment and methylation metrics. This app supports libraries prepared using TruSeq DNA Methylation and TruSeq Methyl Capture Library Prep Kits.	Learn more
 MethylKit	MethylKit provides DNA methylation analysis and annotation from high-throughput bisulfite sequencing. It calculates basic statistics such as methylation statistics, as well as differential methylation regions between two conditions, eg, experiment against reference.	Learn more
 DRAGEN Methylation Pipeline	The DRAGEN Methylation Pipeline rapidly analyzes whole-genome and targeted bisulfite DNA sequence data. It performs alignment and methyl calling, and calculates alignment and methylation metrics. This app supports libraries prepared using TruSeq DNA Methylation and TruSeq Methyl Capture Library Prep Kits.	Learn more

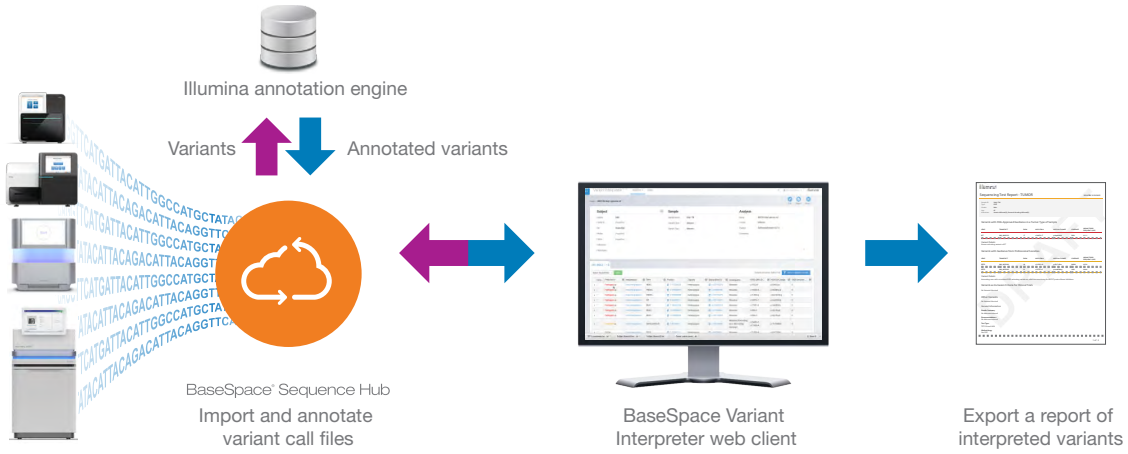


Figure 7: BaseSpace Variant Interpreter—BaseSpace Variant Interpreter is a powerful reporting solution for analyzing and interpreting variant data. This solution aggregates information from a collection of databases to streamline annotation. It also provides flexible filtering options for analyzing variant data and tools to enable classification and reporting of biologically relevant variants.

Bioinformatics solutions for bulk gene regulation studies

Gene regulation is the process of how a given gene is turned on or off in a biological process. Genetic characterization of gene regulation studies can involve detecting and characterizing methylation patterns across the genome, identifying DNA-protein interactions, evaluating regions of open chromatin, and more. For methylation analysis, output files include methylation stats plots, methylation correlation plots, differential methylation summary tables and regions, methylation stats summaries. For DNA–protein analysis, outputs include annotation, peak, and motif files that can be visualized in the Interactive Annotated Peak/Motif Explorer, and alignment files (in BAM file format). For open-chromatin analysis by ATAC-Seq, outputs include results of peak calling, peak annotation, peak differential analysis, nucleosome positioning, and more.[28]

Methylation sequencing

Methylation sequencing data can be analyzed with various BaseSpace apps, including the MethylSeq and MethylKit Apps (Table 9). The MethylSeq App uses Bismark[29] and Bowtie2[30] to map bisulfite-treated sequencing reads to the genome of interest and performs methylation calls. The MethylKit App analyzes sequencing data for differences in methylation between samples. BaseSpace MethylSeq and MethylKit apps support targeted data and common epigenomics analysis tasks such as methylation calling, analysis of differential methylation between samples, and categorization of significant methylation regions. BaseSpace MethylSeq and MethylKit apps are designed for push-button ease of use for any researcher, regardless of bioinformatics experience.

ChIP-Seq

ChIP-Seq data can be analyzed with the ChIPSeq BaseSpace App. It uses Model-based Analysis of ChIP-Seq (MACS)[31] to identify enriched regions pulled down by chromatin immunoprecipitation and HOMER[32] to discover motifs within these regions. The raw outputs of these tools are made available for download and presented in an interactive table. A separate analysis is run for each sample group. A sample group consists of a treatment sample (in which the pull-down was performed) and a control, both of which may be split across multiple replicate FASTQ files.

ATAC-Seq

After initial processing and alignment, ATAC-Seq data can be analyzed using various commercial and open-source algorithms and software programs, including ENCODE ATAC-Seq, chromVAR, and Partek (Table 10). These tools primarily function during peak calling to identify accessible regions of chromatin. Peak calling in ATAC-Seq analysis can use MACS, designed to identify transcription factor binding sites for ChIP-Seq analysis. Outputs include plain text browser extensible data (BED) files with genomic coordinates for called peaks and associated statistics (fold change, p-value, q-value, etc.).

Learn more about ATAC-Seq data analysis pipelines at yiweiniu.github.io/blog/2019/03/ATAC-seq-data-analysis-from-FASTQ-to-peaks/

Table 10: ATAC-Seq analysis tools

Software program	Description	Link
ENCODE ATAC-Seq pipeline	Open-source pipeline designed for automated analysis of ATAC-Seq data, from FASTQ files to peak calling.	Learn more
chromVAR	Open-source R package for analysis of variations in chromatin accessibility to identify associated motifs or genomic annotations.	Learn more
Partek	Commercially available statistical analysis software products for ATAC-Seq data analysis in a user-friendly interface with guided workflows.	Learn more

Summary

To attain biological meaning out of gene expression and regulation NGS experiments, easy-to-use bioinformatics tools are critical and more accessible than ever before. These bioinformatics techniques can assist with providing relevant data and answer numerous questions about gene expression and regulation in many research settings. When coupled with NGS instruments from Illumina, these data analysis tools enable researchers to explore gene expression and regulation at any level of cellular resolution.

Chapter 4: Bioinformatics pipelines for single-cell analyses

Generally, single-cell sequencing data analysis workflows, such as those for scRNA-Seq, scATAC-Seq, and single-cell proteomics, are similar to their complementary bulk protocols. However, single-cell sequencing data analysis requires specialized bioinformatics solutions and novel methods, as the rationale behind the design of bulk sequencing data analysis often are not valid for single-cell data.[33-35]

Challenges with analyzing scRNA-Seq data

Analysis of scRNA-Seq data presents several challenges not applicable to bulk sequencing data analysis. With advances in microfluidics technology enabling high-throughput single-cell isolation, the increasing numbers of cells analyzed translates into significantly more data points that require scalable analysis methods. Inherent to scRNA-Seq data is the increased variability, as compared to bulk RNA-Seq, reflective of the increase in biological complexity. Further complicating scRNA-Seq data analysis are observed zeroes or "dropouts," genes for which no UMIs or sequencing reads map to a particular cell. This phenomenon can result from two possible causes: either the gene is expressed but is not detected (a technical issue) or the gene truly is not expressed by that particular cell. An added layer of complexity is the reality of the temporal nature of gene expression, ie, cells may exhibit high transcriptional activity only at certain times or under certain conditions. These factors and more must be accounted for during primary, secondary, and tertiary analysis of scRNA-Seq data.[36-41]

Primary and secondary analysis of single-cell sequencing data

Primary analysis for single-cell sequencing data proceeds in the same fashion as for bulk sequencing, consisting primarily of file conversion from the BCL to FASTQ format. After conversion, secondary analysis can proceed.

Demultiplexing

Demultiplexing is an important step in single-cell sequencing data analysis. Whereas demultiplexing in bulk sequencing involves separating reads from pooled libraries into individual libraries, in single-cell sequencing reads from pooled samples are separated into individual cells based on cell barcodes added during cell isolation (Figure 8). Many single-cell analysis platforms include demultiplexing capabilities; however, if needed, open-source pipelines like zUMIs are available.

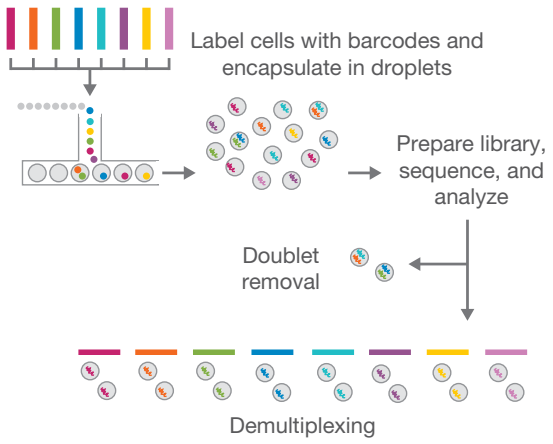


Figure 8: Demultiplexing in single-cell sequencing—Cell barcodes added during isolation are used to parse sequencing reads to individual cells, as well as remove cell doublets during secondary analysis.

QC metrics

Before downstream analysis, several QC metrics can be employed to help determine the quality of a single-cell sequencing data set. These typically include estimated cell counts, intergenic/intronic/exonic content, fraction of reads in cells, expected library size, and number of expressed genes. These metrics cannot be assessed by traditional quantification methods such as fluorometry or qPCR. To maximize the efficiency of high-throughput single-cell experiments, such as cell atlas studies or when combining multiple single-cell libraries, sequencing first at shallow depths on the iSeq 100 System enables characterization of key metrics and subsequent rebalancing before a high-depth NGS run. Library QC leads to more consistent results, which can simplify data analysis and interpretation.

Read the *QC and rebalancing of single-cell gene expression libraries using the iSeq 100 System* application note

Tertiary analysis of single-cell sequencing data

Generally, the analysis pipeline for single-cell sequencing data is to align reads, generate feature-barcode matrices, and perform tertiary analysis. Dimensionality reduction enables clustering of data points and is an important step in the analysis of scRNA-Seq, sc-ATAC-Seq, and single-cell protein profiling.

Dimensionality reduction algorithms

The ability to analyze thousands of data points across hundreds to thousands of cells simultaneously results in single-cell sequencing data having a high dimensionality.[39] Typically, high dimensional data sets undergo dimensionality reduction to ease the computational burden on downstream analysis, reduce noise in the data, and enable data visualization by capturing the underlying structure in the data set in two or three dimensions.[43] There are several algorithms available to perform dimensionality reduction.

PCA

Principal component analysis (PCA) is a commonly used algorithm that performs linear dimensional reduction, projecting data to a lower number of independent dimensions for maximal variance capture (Figure 9A). PCA assumes data are approximately normally distributed, which may not always apply to single-cell sequencing data.[39]

t-SNE

t-distributed stochastic neighborhood embedding (t-SNE) is a nonlinear computational technique for dimensionality reduction. t-SNE makes multidimensional data understandable by mathematically reducing the number of dimensions into a two- or three-dimensional representation and is commonly used to visualize subpopulations with single-cell sequencing data. However, t-SNE requires long computing times, is limited in its ability to represent very large data sets, and does not preserve global structure. This means that while distances between data points within a cluster are meaningful and informative, intercluster distances are not (Figure 9B).[43]

UMAP

Uniform manifold approximation and projection (UMAP) is a nonlinear technique for dimensionality reduction of any type of high dimensional data that can be applied to biological and single-cell sequencing data sets. UMAP preserves local and global structure within large data sets, with relatively fast compute times.[44]

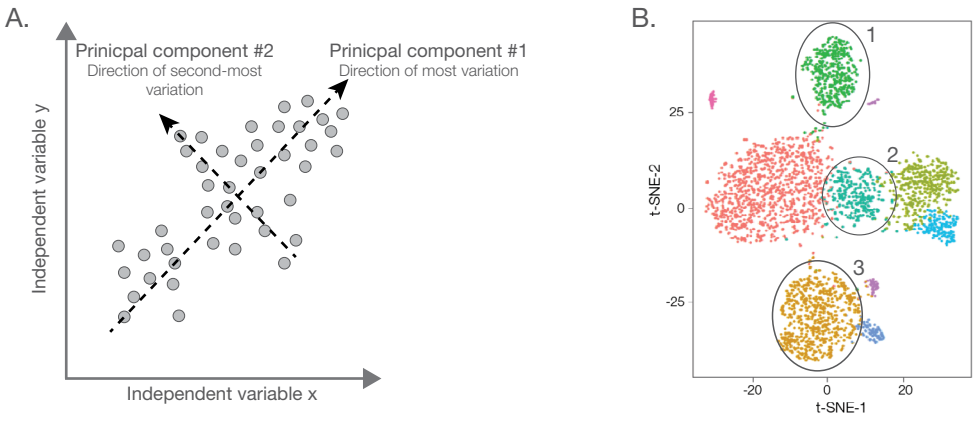


Figure 9: Clustering with PCA and t-SNE—(A) PCA projects high dimensional data to a lower number of dimensions for maximal variance capture, but is restricted to linear dimensions and assumes normally distributed data. (B) t-SNE reduces dimensionality, plotting data points on a two- or three-dimensional plot. Importantly, while clusters denote similarity between data points therein, the distance between clusters do not indicate similarity, ie, clusters 1 and 2 are not necessarily more similar than clusters 1 and 3.

Downstream tertiary analysis of single-cell sequencing data

Additional downstream analysis steps depend on the specific method, for example, scRNA-Seq analysis features differential gene expression analysis.[42] As a complement to gene expression studies with scRNA-Seq, epigenetic studies via scATAC-Seq can elucidate gene regulation.[47] Analyzing scATAC-seq experiments can be challenging because the datasets tend to be large, sparse, and binary, but several analysis solutions are available to gain biological insights from chromatin accessibility data.[28]

Single-cell proteome analysis enabled via NGS as the readout for protein expression within single cells involves the three major approaches: AbSeq,[4] CITE-Seq,[5] and REAP-Seq.[6] Analysis of the resulting data sets can be challenging, given its multimodal nature; however, antibody-tag sequences (ADTs) can be quantified in much the same manner as UMI counts for transcript quantification. Analysis of these data sets is facilitated by updates to existing tools for scRNA-Seq analysis to support multimodal data and newly developed tools.

There are many options for single-cell tertiary analysis tools, including open-source analysis tools developed by academic labs in popular programming languages like R and Python, 'plug-and-play' packages that allow researchers to use preconfigured analysis workflows, and commercial offerings. The tools chosen will depend on the research goals and experimental objectives.

Open-source bioinformatics tools (freeware)

Seurat

Seurat is an R-based scRNA-Seq analysis software designed to assess cellular heterogeneity using normalization, dimensionality reduction approaches, plots, heat maps, and data integration tools.[35] Seurat uses dimensionality reduction to cluster cells into groups that correspond to particular cell states or types with characteristic features.

Seurat users can integrate single-cell data sets generated across different conditions, technologies, or species. The application can also explore multimodal data, such as scATAC-Seq and single-cell proteomics data in combination with scRNA-Seq data.

Learn more at satijalab.org/seurat/v3.0/integration.html

Monocle

Monocle is an R-based scRNA-Seq analysis software designed to determine cell developmental trajectory. Monocle is ideal for experiments where there are known beginning and terminal cell states. The software uses machine learning techniques to order individual cells along differentiation patterns, clusters cells using t-SNE, and performs differential gene expression analysis

Learn more at cole-trapnell-lab.github.io/monocle-release/

Read an interview with Cole Trapnell about advances in single-cell analysis at www.illumina.com/science/customer-stories/icommunity-customer-interviews-case-studies/trapnell-uwash-interview-single-cell.html

velocyto

RNA velocity describes the rate of change in gene expression for a particular gene at a specific point in time based on the ratio of spliced and unspliced mRNA. scRNA-Seq can determine RNA velocities within single cells to predict trajectories during differentiation and analyze cell lineages over the course of hours to days. The velocyto software package enables analysis of gene expression dynamics in scRNA-Seq data to estimate RNA velocity.[45,46]

Learn more at velocyto.org/

chromVAR

chromVAR is an open-source R package for analyzing variations in chromatin accessibility in bulk or scATAC-Seq data to identify associated motifs or genomic annotations. The software enables identification of known and *de novo* sequence motifs associated with chromatin accessibility.[48]

Learn more at github.com/GreenleafLab/chromVAR

Human Cell Atlas Data Coordination Platform

The Human Cell Atlas Data Coordination Platform (HCA DCP) is an open-source software tool for data coordination 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. Single-cell data can be submitted to the HCA DCP by labs and researchers around the globe.

Learn more at data.humancellatlas.org/

CITE-Seq-Count and CITEFuse

Bioinformatics tools such as CITE-seq-Count and CiteFuse[49] can be applied to various single-cell protein profiling methods to support data analysis and visualization.

Learn more at cite-seq.com/

Platform-specific commercially available bioinformatics tools

Various software tools developed by commercial providers support specific platforms for single-cell isolation and analysis, including 10x Genomics, MissionBio, and others (Table 11).

Table 11: Platform specific software for single-cell sequencing data analysis

	Description	Sequencing data type	Link
10x Loupe Browser	The 10x Loupe Browser is a desktop application that allows researchers to visualize and analyze single-cell sequencing data, including 10x Chromium scRNA-Seq and sc-ATAC-Seq data. The software enables users to find significant genes, cell types, and substructure within scRNA-Seq data quickly and interactively. The 10x Loupe Browser analyzes sc-ATAC-Seq data to determine differential chromatin accessibility, find significant regulatory features, distinguish transcription factor motifs, and more.	scRNA-Seq sc-ATAC-Seq	Learn more
10x Cell Ranger ATAC	Cell Ranger ATAC is a set of analysis pipelines designed to process 10x Chromium Single Cell ATAC data. The software performs primary analysis, including file conversion, demultiplexing, read filtering, and alignment; secondary analysis, including identification of transposase cut sites, detection of accessible chromatin peaks, cell calling, and count matrix generation for peaks and transcription factors; and tertiary analysis, including dimensionality reduction, cell clustering, and clustering of regions of differential accessibility.	sc-ATAC-Seq	Learn more
Tapestri Insight	Tapestri Insight is a software solution for analyzing Tapestri single-cell DNA sequencing data that includes sequence import, data analysis, and visualization capabilities. The software enables variant identification, including SNVs and copy number variants (CNVs), at the clonal and subclonal level.	single-cell DNA sequencing	Learn more
Bio-Rad SureCell ATAC-Seq Analysis Toolkit	The SureCell ATAC-Seq Analysis Toolkit is used with the SureCell ATAC-Seq Library Prep Kit to enable epigenetic analysis of single cells genome-wide. The software can estimate gain and loss of chromatin accessibility within peaks, cluster scATAC-Seq profiles, characterize of sequence motifs associated with gene expression, and identify cis- and trans-acting elements that are the source of diverse cellular phenotypes.	sc-ATAC-Seq	Learn more

Platform-agnostic commercially available bioinformatics tools

SeqGeq™ v1.6 Software

SeqGeq Software, developed by the makers of FlowJo Software, is a platform-agnostic desktop bioinformatics platform designed for the analysis of single-cell experiments. It includes a wide suite of informatics features including: V(D)J analysis, Seurat clustering, Monocle trajectory inference, and more. All of these tools are designed to be compatible with data from any instrument or sequencing pipeline.

SeqGeq allows researchers to perform advanced analysis, data exploration, and visualization with an easy to use drag-and-drop interface familiar to anyone who uses FlowJo. It generates quality figures that are easily shared for publication and collaboration (Figure 10) and integrates with BaseSpace Sequence Hub directly to complete the data analysis workflow.

Learn more at www.flowjo.com/solutions/seqgeq

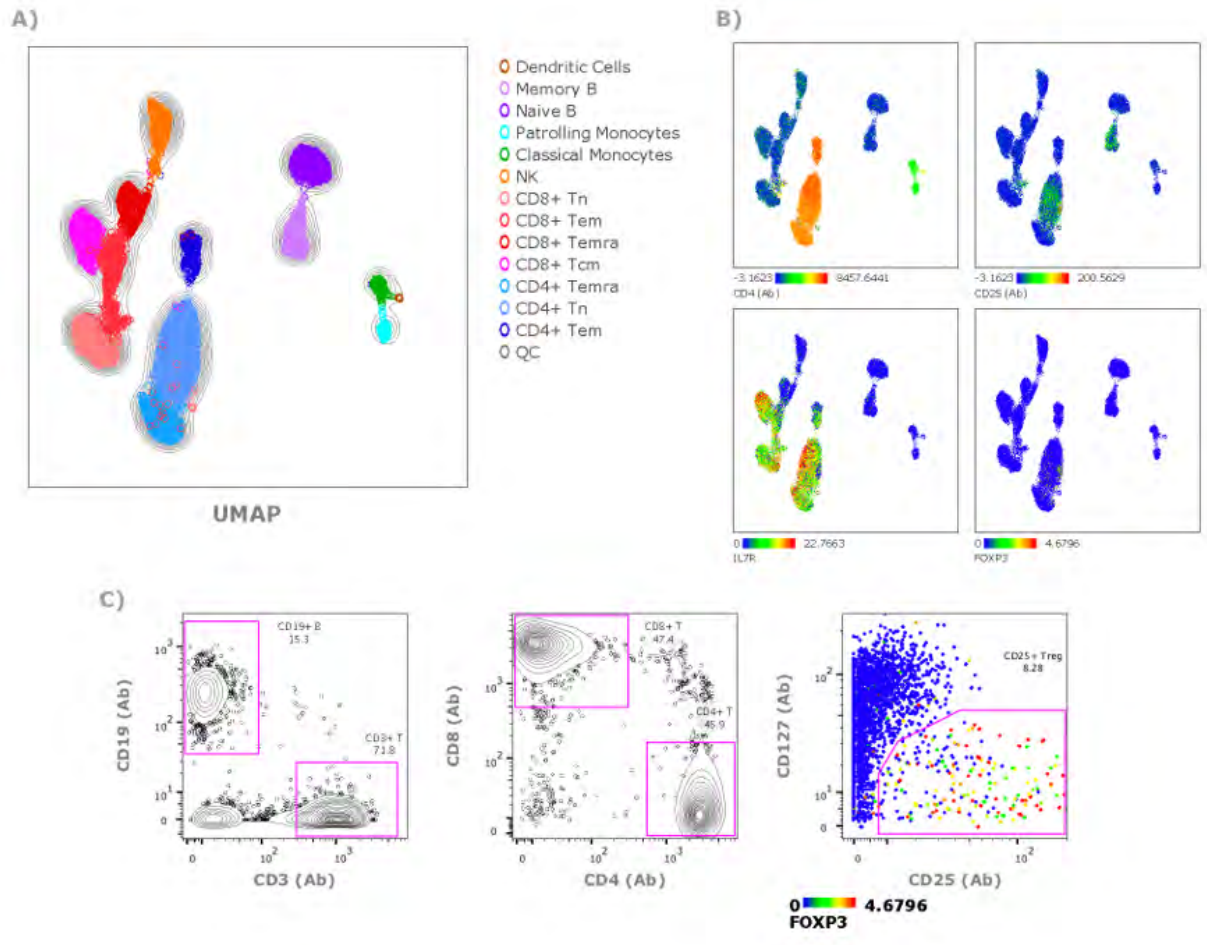


Figure 10: Standard analysis of peripheral blood samples in SeqGeq—(A) Dimensionality reduction, clustering, and annotations generated through easy to use GUI, (B) cells heat mapped by gene and antibody expression, (C) traditional gating structure leading to a Treg population of interest with manual gating analysis. Data was generated and provided by BD Biosciences using SeqGeq Software.

Partek Flow

With robust statistics and interactive visualizations, Partek Flow allows researchers to quickly and reliably discover biological meaning in bulk and single-cell genomic data, without the need for advanced bioinformatics skills. Partek has been used by researchers around the world for data analysis, being cited in over 8000 peer-reviewed publications. Partek is compatible with mutliple single-cell applications, including scRNA-Seq, CITE-Seq, cell hashing, spatial transcriptomics, and more (Figure 11). Analysis capabilities of Partek include:

- Classify cells into known and novel cell types
- Discover biomarkers that define a cell population
- Find differentially expressed genes, proteins, and pathways
- Compare cell type populations between experimental groups/phenotypes
- Integrate gene and protein expression data from multi-omics experiments

Learn more at www.partek.com/partek-flow/

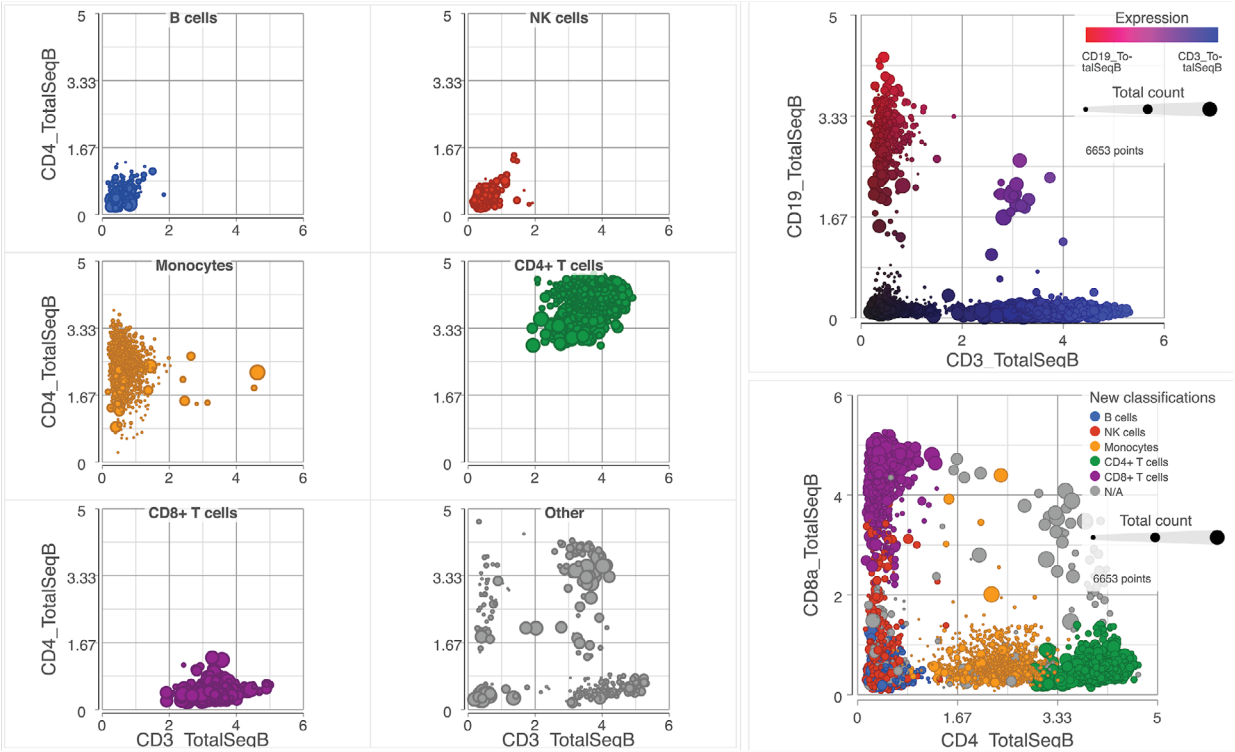


Figure 11: Analysis of single-cell samples in Partek Flow—Partek Flow offers robust statistical algorithms, information-rich visualizations, and cutting edge genomic tools. Starting from from a BCL, FASTQ, BAM, FCS, H5, TXT, CSV, or CBCL file, Partek Flow enables rich and interactive visualizations. Data was generated and provided by Partek using Partek Flow software.

Summary

In contrast to bulk sequencing, single-cell sequencing studies require novel methods for data analysis. A wide variety of open-source and commercially available bioinformatics tools enable analysis, visualization, and interpretation for scRNA-Seq, scATAC-Seq, and protein profiling studies. The ability to derive meaningful insights from these methods is key to understanding the complexity of cellular and molecular biology.

Conclusion

NGS-based gene expression and regulation studies with bulk or single-cell samples have great potential to elucidate complex biological systems. With increasing options for investigating gene expression and regulation, there has been a remarkable diversification of bioinformatics pipelines, each with inherent strengths and weaknesses. Careful pipeline design and optimization throughout the analysis workflow is crucial for success.

Here, we have presented every step of the NGS workflow, with a focus on the computational analysis pipelines available for analyzing bulk and single-cell data. Important considerations and potential challenges have been discussed, commercial offerings presented, and advice given for executing an analysis workflow for successful NGS-based gene expression and regulation studies.

References

1. Ozsolak F, Milos PM. RNA Sequencing: advances, challenges and opportunities. *Nat Rev Genet*. 2011;12:87–98.
2. Wang W, Niu Z, Wang Y, et al. Comparative transcriptome analysis of atrial septal defect identifies dysregulated genes during heart septum morphogenesis. *Gene*. 2016;575:303–312.
3. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29–39.
4. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci Reports*. 2017;7:44447.
5. Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14:865–868.
6. Peterson VM, Zhang KX, Kumar N, et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotech*. 2017;35:936–939.
7. Illumina (2018). iSeq 100 Sequencing System specification sheet. Accessed July 16, 2020.
8. Illumina (2019). MiniSeq Sequencing System specification sheet. Accessed July 16, 2020.
9. Illumina (2018). MiSeq System specification sheet. Accessed July 16, 2020.
10. Illumina (2019). NextSeq 550 Sequencing System specification sheet. Accessed July 16, 2020.
11. Illumina (2020). NextSeq 1000 and NextSeq 2000 Sequencing Systems specification sheet. Accessed July 16, 2020.
12. Illumina (2019). NovaSeq 6000 Sequencing System specification sheet. Accessed July 16, 2020.
13. Sims D, Sudbery I, Ilot NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121–132.
14. Corley SM, MacKenzie KL, Beverdam A, Roddam LF, Wilkins MR. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC Genomics*. 2017;18:399.
15. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*. 2013;8(10):e77910.
16. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
18. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
19. Agilent Technologies. (2016). RNA Integrity Number (RIN) – Standardization of RNA Quality Control Publication. PN-5989-1165EN. Accessed August 4, 2020.
20. Wong KS, Pang H. Simplifying HT RNA Quality & Quantity Analysis. *Genet Eng & Biotech News*. 2013;33(2):4688.
21. Sheng Q, Vickers K, Wang J, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. *Brief Func Genomics*. 2017;16(4):194–204.
22. Guo Y, Long J, He J, et al. Exome sequencing generates high quality data in non-target regions. *BMC Genomics*. 2012;13:194.
23. Illumina (2020). BaseSpace Sequence Hub data sheet. Accessed August 4, 2020.
24. Illumina (2019). Illumina DRAGEN Bio-IT Platform data sheet. Accessed August 4, 2020.
25. Illumina (2018). BaseSpace Correlation Engine data sheet. Accessed August 4, 2020.
26. Sahraeian SME, Mohiyuddin M, Sebra R, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun*. 2017;8(1):59.
27. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562–578.
28. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol*. 2020;21(1):22.
29. Krueger F and Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011; 27:1571–1572.
30. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–359.
31. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
32. Heinz S, Benner C, Spann N, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 2010 May 28;38(4):576-589.
33. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50(8):96.
34. Eberwine J, Sul JY, Bartfai, T, Kim J. The promise of single-cell sequencing. *Nat Methods*. 2014;11(1):25–27.
35. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. 2017;9(1):75.
36. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019;15(6):e8746.
37. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21(31):doi.org/10.1186/s13059-020-1926-6.
38. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131(4):281–285.
39. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet*. 2019;10:317.
40. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32(9):896–902.
41. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381–386.
42. Gao M, Ling M, Tang X, et al. Comparison of High-Throughput Single-Cell RNA Sequencing Data Processing Pipelines. *bioRxiv*. 2020;02.09.940221.
43. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. *Mol Aspects Med*. 2018;59:114–122.
44. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2019;37:38–44.
45. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. *Nature*. 2018;60(7719):494–498.
46. Bergen V, Lange M, Peidli S, Wolf A, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020; doi: 10.1038/s41587-020-0591-3.
47. Buenrostro J, Wu B, Litzenburger U, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015;523(7561)486–490.
48. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*. 2017;14:975–978.
49. Kim HJ, Lin Y, Geddes TA, Hwa Yang JY, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*. 2020;36(14):4137–4143.

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

illumina®