Unlock insights from gene expression and regulation studies

Accurate, versatile, and integrated data analysis solutions for high-impact research with Illumina Connected Software

illumina®

M-GL-00060 v1.0

Table of contents

Using the power of NGS to study gene expression and regulation

Genes encode for a vast array of protein- and nonprotein-coding RNA elements that drive essential biological processes, such as development, growth, and differentiation. Understanding the transcriptome—the complete set of RNA transcripts within a cell—is essential for interpreting the functional elements of the genome and providing valuable insights into health and disease states.

Next-generation sequencing (NGS) has revolutionized the study of the transcriptome by enabling simultaneous profiling of thousands of genes on a genome-wide scale. NGS-based RNA sequencing (RNA-Seq) is a highly sensitive and accurate tool that delivers a high-resolution, base-by-base view of RNA activity for measuring gene expression across the transcriptome. This method can uncover previously undetected changes occurring in disease states, in response to therapeutics, under different environmental conditions, and across a broad range of other study designs. RNA-Seq captures important, but low-frequency molecular events that might be missed using targeted methods, making it an excellent tool for transcript and variant discovery studies. Because RNA-Seq does not require predesigned probes, the data obtained are unbiased, allowing for hypothesis-free experimental design.

Additionally, advances in NGS methods have facilitated the development of sophisticated methods for studying gene regulation. For example, methylation sequencing (Methyl-Seq) can be used to study cytosine methylation patters across the genome, while chromatin immunoprecipitation with sequencing (ChIP-Seq) identifies genome-wide DNA binding sites for transcription factors and detects DNA–protein interactions. Another approach for studying gene regulation is the Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-Seq), which assesses chromatin accessibility. These powerful methods provide valuable insight into the cellular mechanisms that control gene expression.

Key terms

Gene expression is the process by which DNA is translated into functional, biologically active units. Studying differential gene expression provides crucial insight into cellular physiology and how dysregulation of these processes leads to the development of disease.

Gene regulation is a dynamic process that orchestrates the timing, location, and amount of gene expression. Key factors that influence gene expression include structural organization, or DNA packaging, and chemical modification of DNA. Uncovering patterns of gene regulation provides an additional dimension of understanding how gene expression is controlled in response to environmental stimuli in health and disease states.

The **transcriptome** is the complete set of RNA transcripts in a cell and their quantities under specific conditions.

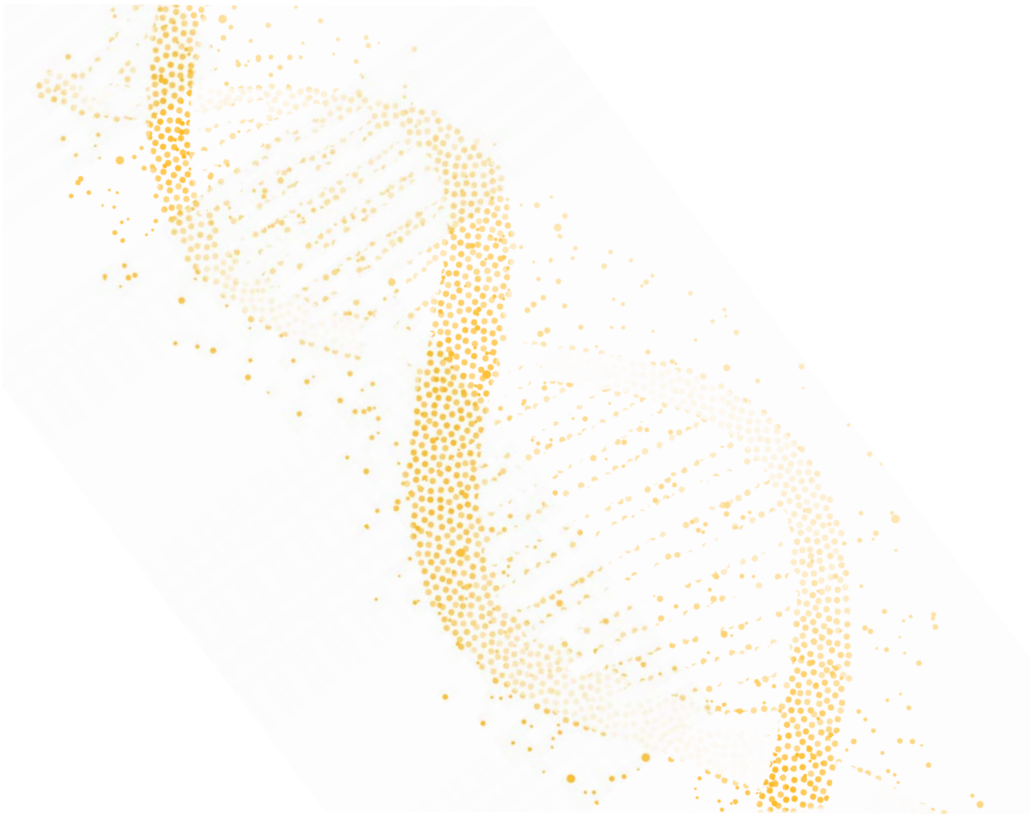
Analyzing and interpreting the large amounts of data generated by these NGS approaches create significant bottlenecks for laboratories. Illumina understands that with the rapid advances in the gene expression and regulation space, there is a need for streamlined sample-to-insight workflows. To address this challenge, researchers can now access the most accurate analysis, versatile, secure, and integrated NGS software suite with Illumina Connected Software. These accessible solutions support researchers across the entire NGS workflow, from lab and sample management, through insight generation.

This ebook outlines the NGS analysis workflow and provides an overview of Illumina Connected Software solutions available for analyzing gene expression and regulation data. Important considerations and best practices for designing and executing a successful NGS-based gene expression and regulation study are discussed, along with specific experimental use cases.

Learn more

NGS accelerates transcriptomics and epigenomics research

RNA-Seq applications



Important considerations for experimental design

Gene expression and regulation data are characterized by high variability and dimensionality. Careful study design that accounts for these factors is crucial for robust statistical analysis and successful transcriptomics and epigenomics experiments. Illumina recommends consulting the primary literature for your field and organism for the most up-to-date guidance on experiment design.

Key considerations while designing gene expression and regulation studies

Aim to minimize variability by identifying all possible sources of variation in samples

Include at least three biological replicates for each sample type

Perform pilot studies to identify how many biological replicates and read depth are needed for sufficient statistical power

Determining appropriate read length and depth

Read depth: For RNA-Seq experiments, the number of reads per sample, or read depth, is typically used instead of coverage. Increased read depth is required for detecting low-abundance genes.

Read length: Paired-end RNA-Seq enables discovery applications such as detecting gene fusions and characterizing novel splice isoforms. For gene expression profiling, single-end sequencing may be sufficient; however, paired-end reads can enable more accurate read alignment and the ability to detect insertion–deletion variants, which is not possible with single-read data.

Learn more

RNA-Seq Workflow Guide

Considerations for RNA-Seq read length and coverage

New to NGS?

Illumina implementation partners can help design and build your informatics solutions to get you up to speed faster. Contact implementation_partners@illumina.com for more information.

Steps in a typical NGS workflow

NGS workflows start with nucleic acid isolation followed by library preparation. Libraries are sequenced on Illumina sequencing systems, designed to support a wide range of applications and throughputs. Generated data are then analyzed to gain insights (Figure 1).

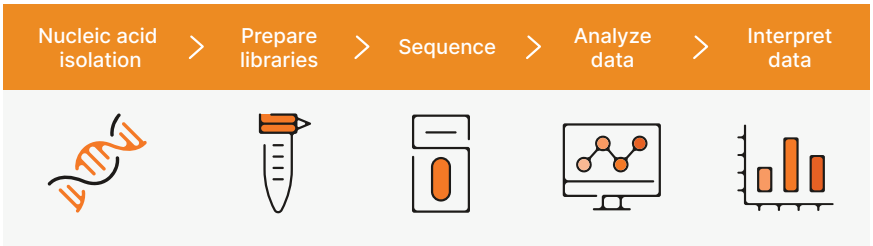


Figure 1: A typical NGS workflow.

Step 1: Library preparation

A library is a collection of similarly sized nucleic acid fragments with known adapter DNA sequences attached to the 5' and 3' ends. Library preparation is a critical step in the NGS workflow as it prepares nucleic acid samples to be compatible with Illumina sequencing systems. Sequencing libraries for RNA samples are typically created by first reverse transcribing the RNA to complementary DNA (cDNA) and then adding specialized oligonucleotide adapters to both ends of the cDNA. Prepared libraries are hybridized to special slides known as flow cells. In the Illumina sequencing workflow, the adapters contain complementary sequences that allow the cDNA fragments to hybridize to the flow cell. Fragments can then be amplified and purified.

Multiple libraries can be pooled together and sequenced in the same run, using a process known as multiplexing. During adapter ligation, unique index sequences, or 'barcodes', are added to each library. These barcodes are used to distinguish between the libraries during data analysis.

Learn more

RNA library preparation

Illumina library preparation products for gene expression and regulation

Illumina Stranded Total RNA Prep with Ribo-Zero™ Plus: Enables comprehensive transcriptome analysis, accurate measurement of gene and transcript abundance, and detection of known and novel coding features and multiple forms of noncoding RNA.

Illumina Stranded mRNA Prep: Quantifies gene expression, identifies known and novel isoforms in the coding transcriptome, detects gene fusions, and measures allele-specific expression.

Illumina RNA Prep with Enrichment: Enables cost-effective RNA exome analysis using sequence-specific capture of the coding regions of the transcriptome. Ideal for formalin-fixed paraffin-embedded (FFPE) samples.

TruSeq™ Small RNA Library Preparation Kit: Offers a simple, cost-effective solution for preparing microRNA (miRNA) and small RNA sequencing libraries from total RNA.

Illumina Tagment DNA TDE1 Enzyme and Buffer Kit: Supports ATAC-Seq studies for genome-wide chromatin accessibility profiling.

Learn more

Illumina library preparation kits

Step 2: Sequencing

Flow cells are inserted into the Illumina sequencing system, where sequencing takes place. Illumina uses proven sequencing-by-synthesis (SBS) and powerful XLEAP-SBS™ chemistry to detect single bases as they are incorporated into growing DNA strands. Several sequencing platforms are available to support a broad range of throughputs and applications. Regardless of your research questions, the flexibility of Illumina sequencing systems can help you get to those answers faster, with simple push-button workflows (Figure 2).

	Benchtop systems				Production-scale systems		
							
Application	iSeq 100	MiniSeq	MiSeq Series*	NextSeq 550 Series*	NextSeq 1000 & NextSeq 2000	NovaSeq™ 6000 Series*	NovaSeq X Series
Targeted gene expression profiling	●	●	●	●	●	●	●
miRNA and small RNA analysis	●	●	●	●	●	●	●
Transcriptome sequencing (Total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●	●	●
DNA–protein interactions (ChIP-Seq)			●	●	●	●	●
Chromatin accessibility (ATAC-Seq)					●	●	●
Methylation sequencing				●	●	●	●
Single-cell analysis (scRNA-Seq, sc-ATAC-Seq)				●	●	●	●
Specifications							
DRAGEN™ secondary analysis on board	No	No	No	No	Yes	No	Yes
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp	2 × 150 bp
Maximum reads per run	4M	25M	25M	400M	1.2B	20B	52B

*Diagnostic options available

Figure 2: Illumina sequencing systems at a glance.

Learn more

Illumina sequencing platforms

Step 3: Data analysis

Analyzing sequencing data involves primary, secondary, and tertiary analysis steps. **Primary analysis,** or base calling, is completed automatically on Illumina sequencing systems. Base calls are signals generated by the sequencing system used to infer the order of the nucleotides in the sample that was sequenced. **Secondary analysis** steps involve demultiplexing, mapping, alignment, and gene expression profiling. This includes differential expression, gene fusion detection, and RNA variant calling. Finally, data visualization and biological interpretation are completed during **tertiary analysis**—converting data into insights. Several approaches and software options are available for each analysis step depending on specific research objectives.

Learn more

Data analysis solutions

Summary

In addition to optimized study design, data analysis, visualization, and interpretation are critical steps for maximizing insights from gene expression and regulation studies. User-friendly, scalable software solutions from Illumina empower researchers to gain a deeper understanding of transcriptomic and epigenomic data, without the need for extensive bioinformatics expertise. These powerful Illumina Connected Software solutions will be discussed in depth in the following sections.

Key terms

Demultiplexing
Barcodes are used to identify which sequence comes from which cells in multiplexed samples that are sequenced together.

Alignment
Sequences are aligned with reference genomes for comparison.

Expression profiling
Data sets are analyzed for gene expression, transcript quantification, or differential expression analysis.

Variant annotation
Functional information is assigned to RNA variants and gene fusions.

Informatics workflow for analyzing gene expression and regulation data

Until recently, analysis and interpretation of the vast amounts of data generated by gene expression and regulation studies required deep bioinformatics expertise. Illumina Connected Software simplifies the process of gaining insights, providing optimized solutions for every step of the NGS workflow—from library preparation to expression profiling, variant interpretation, and visualization. Our integrated software portfolio features secure, versatile solutions with award-winning accuracy and direct integration with Illumina sequencing systems, connecting data across the wet and dry lab. Push-button software tools and apps with intuitive user interfaces support a wide range of applications, including differential gene expression, transcriptome profiling, and more, enabling biologists to maximize the discovery power of their studies.

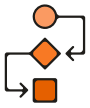
The following section, designed for users who are new to bioinformatics, will delve into the analysis portion of the NGS workflow using Illumina software solutions.

Benefits of Illumina data analysis solutions for RNA-Seq

Accessible to all researchers, regardless of bioinformatics expertise

Support common gene expression research applications

Compatible with all Illumina sequencing systems



Enable streamlined workflows with low manual touchpoints

Steps in the NGS analysis workflow

The NGS data analysis workflow consists of three main steps: primary analysis or base calling, secondary analysis, and tertiary analysis or interpretation (Figure 3).

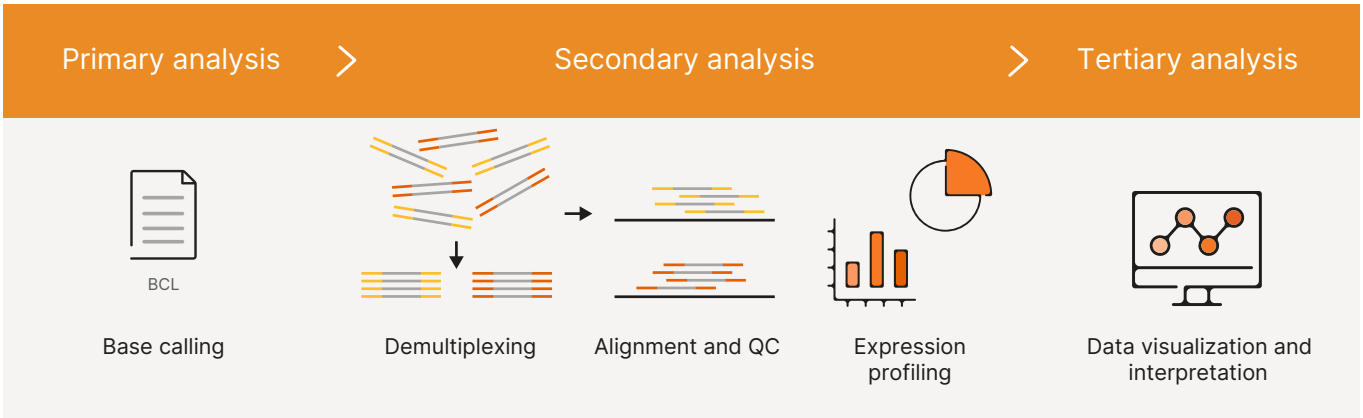


Figure 3: NGS data analysis workflow—NGS data analysis includes three main steps. In primary analysis, raw base call (BCL) files are generated. During secondary analysis, reads are demultiplexed and aligned to a reference genome. The resulting sequence undergoes basic expression profiling. Tertiary analysis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

Primary analysis

Primary data analysis consists of the digitization of genetic information into 'reads' of nucleotide sequences, known as base calling, and scoring base quality. Primary analysis, or Real-Time Analysis (RTA), operates during cycles of sequencing chemistry and imaging, using an on-instrument software to provide base calls and associated quality scores representing the primary structure of DNA or RNA strands. Illumina sequencing systems use built-in RTA software to perform primary data analysis automatically.

Secondary analysis

After sequencing is complete, secondary analysis for genetic characterization and expression profiling begins. BCL files generated during primary analysis are converted into a more versatile, text-based format (FASTQ) (Figure 4). Secondary analysis steps include demultiplexing, mapping, alignment with a reference genome, quality control (QC), gene or transcript quantification, differential expression analysis, gene fusion detection, and RNA and DNA variant calling (Figure 5). Once sequence alignment is complete, the BAM files are analyzed to identify mutations or other features that differ from the reference genome. These variant calls are represented in a text tab-delimited file known as variant call format, or VCF file. Transcript quantification results are reported in <outputPrefix>.quant.sf text files, which can be input for differential gene expression using tools such as tximport and DESeq2.

Figure 4: Example FASTQ read with 50 base calls—FASTQ files are text files containing sequence data with a quality (Phred) score for each base, represented as an ASCII character. The quality score is an integer (Q) which is typically in the range 2–40, but higher and lower values are sometimes used. There are always four lines per read. The first line starts with '@', followed by the label. The second line is the sequence. The third line starts with '+'. In some cases, the '+' line contains a second copy of the label. The fourth line contains the Q scores represented as ASCII characters.

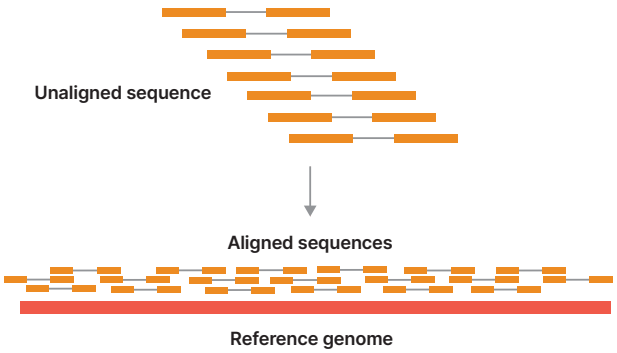


Figure 5: Sequence alignment in secondary analysis—The sequence alignment step maps each base call to the corresponding location on the reference genome or transcriptome. The output of this step is a binary alignment map, or BAM file.

For RNA-Seq data, gene expression analysis includes gene or transcript counts and differential expression data. Data outputs are in a tab-delimited format, such as tab separated value (TSV) files. These files contain columns representing samples, genes, amplicon IDs, raw counts, normalized counts, p-value, fold change, and more.

Transcript quantification (quant.sf) or **gene quantification** (quant.gene.sf) files can be used to assess differential expression between sample groups.

Commonly used file formats for Illumina sequencing data

BCL: Binary base call files that contain raw data generated by Illumina sequencing systems.

FASTQ: Text-based sequencing data file format that stores both raw sequence data and quality scores. FASTQ files are 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 pipelines.

FASTQ.ORA: Lossless compression file format of FASTQ, which reduces the size, time to transfer, and storage cost, without compromising data integrity.

SAM: Sequence alignment map files are a text file format that contains the alignment information of sequences mapped to a reference sequence.

BAM: Binary alignment map files are the output obtained from sequence alignment in binary format. They are smaller and more efficient for software to process than SAM files.

CRAM: Highly compressed alternative to BAM files containing only base calls that differ from the reference.

VCF: Variant call format is a standardized text file format used for storing variant information (single nucleotide polymorphisms (SNPs), indels, fusion genes, and small variants).

Tertiary analysis

The final step of the NGS data analysis workflow is tertiary analysis, which adds biological context to the results generated from secondary analysis. Tertiary analysis of expression profiling leads to powerful insights across multiomic disciplines, facilitating new discoveries into basic biology and disease pathogenesis (Figure 6). Tertiary analysis typically consists of variant annotation, filtering, prioritization, data visualization, and reporting.

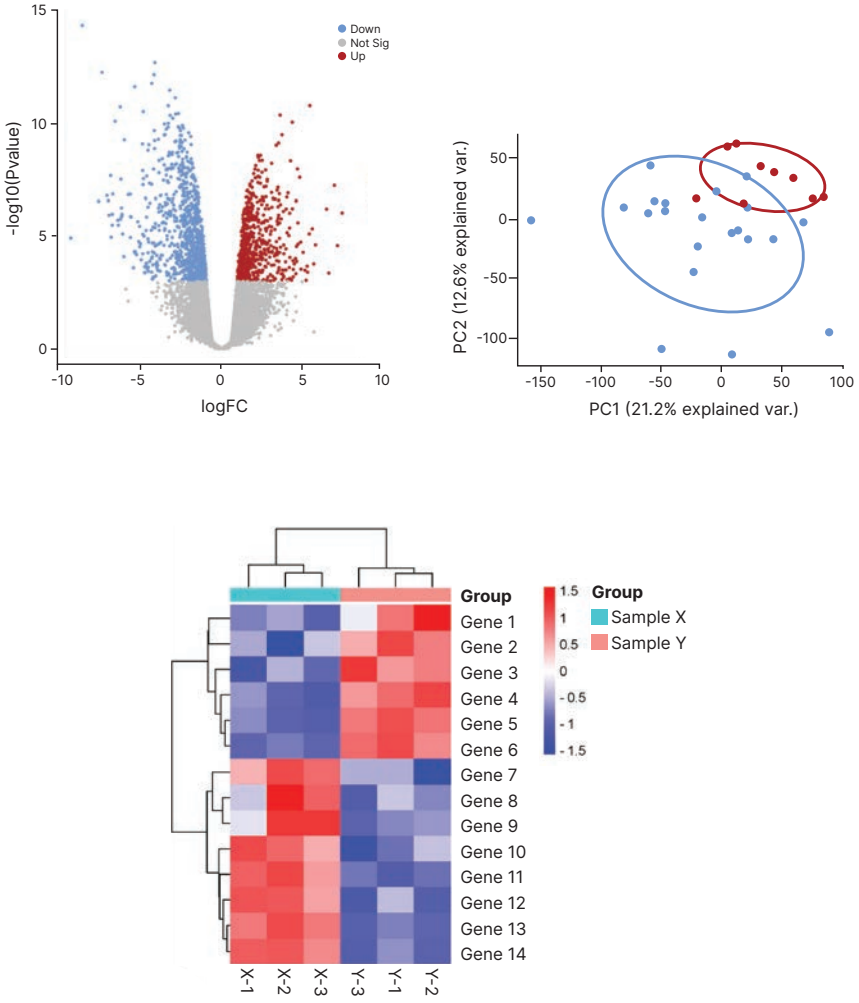


Figure 6: Example outputs from tertiary analysis—Tertiary analysis converts sequencing data into biological insights. Shown here are examples of differential gene expression analysis using (A) volcano plots, (B) principal component analysis, and (C) an expression heat map.

Secondary analysis

Transcript quantification: Quantifying gene expression by counting the number of reads that align to each gene (or read count).

Differential gene expression analysis: Measuring differences in normalized read count data in a group of genes or RNA transcripts to quantify changes in expression levels across experimental groups.

Gene fusion calling: Identifying hybrid genes formed as a result of translocation, deletions, or chromosomal inversions.

Tertiary analysis

Variant annotation: Labeling and assigning genetic characteristics to specific variants.

Filtering: Filtering variants using user-defined gene lists, quality measures, population frequencies and annotation details.

Data visualization: Visualizing reads that support each variant call when reviewing data, including expression heat maps, volacano plots, PCA plots, etc.

Data reporting: Summarizing findings into structured reports that can be streamlined through customized workflows.

Key considerations for planning your data analysis workflow

Careful pipeline design and optimization throughout the analysis workflow is crucial for successfully investigating gene expression and regulation. The choice of software will depend on the specific experimental questions, throughput, and needs of individual labs. It is important to consider what security and compliance requirements are needed, the scalability of analysis pipelines, and user experience. For users who are new to NGS, we recommend collaborating with your core lab partners early in the experimental design process to make sure that your study is optimally designed for robust statistical analysis.

Illumina Connected Software

Illumina Connected Software delivers award-winning accuracy, security, and efficiency for every step of the NGS workflow from sample to insights (Figure 7). After the sequencing run is complete, Illumina Connected Software enables secure data management, and secondary and tertiary analysis. Illumina software solutions meet users where their data is, offering a range of informatics solutions with varying deployment accessibility, such as on board the instrument and in the cloud, for a streamlined sample-to-answer workflow.

The following section will delve deeper into Illumina data analysis and interpretation solutions.

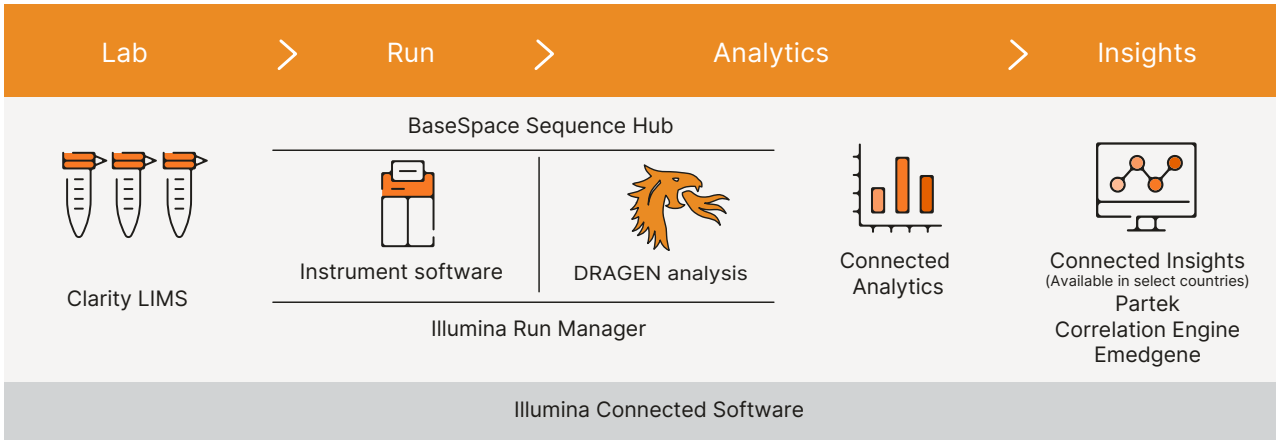


Figure 7: Software for every step of the workflow—Direct instrument integration, auto-launch capabilities, and an interconnected platform enable a high degree of automation, significantly reducing hands-on time by pulling and consolidating data from every step of the workflow.

Learn more

Illumina informatics overview

Lab and run management software

Lab | Run | Analytics | Insight

Clarity LIMS™ software

Clarity LIMS software is a laboratory information management system (LIMS) designed to track samples and manage workflows for optimized laboratory operations (Figure 8). This software powers efficient genomics sample and workflow management, enabling labs to track samples, streamline complex tasks, generate sample sheets, and catch poor-quality samples before running them on the sequencing system. Clarity LIMS software connects with your Illumina sequencing systems with ready-to-use, preconfigured protocols to get you started right away without the need for coding expertise. This flexible system scales with your laboratory needs and can easily accommodate third-party instruments and software as your lab grows.

Clarity LIMS software
Efficient sample tracking and workflow management

Built for: Genomics labs looking for a centralized platform for managing laboratory data, fast-tracking Illumina workflows while reducing manual errors, and meeting compliance standards with industry regulations

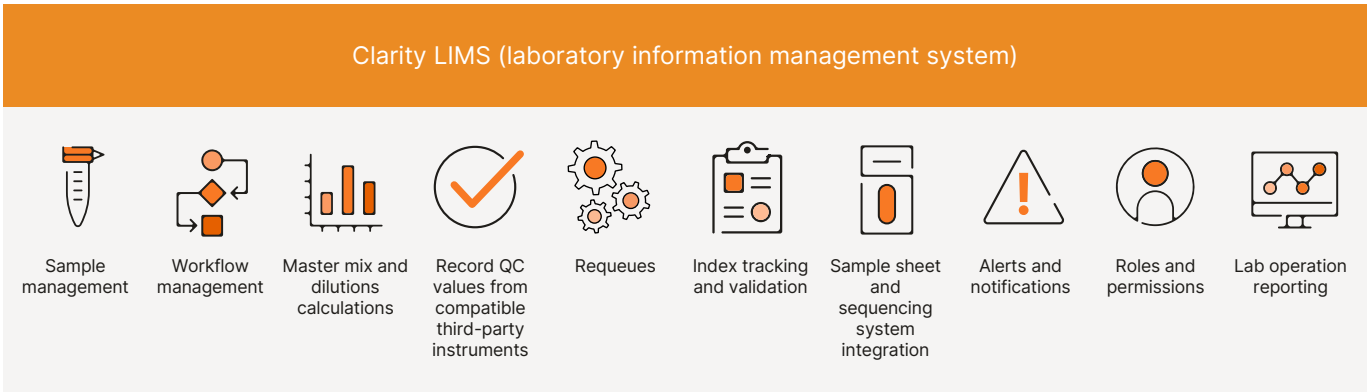


Figure 8: Clarity LIMS software functions.

Learn more

Clarity LIMS software

Run management software

Lab | **Run** | Analytics | Insight

Illumina Run Manager

Illumina Run Manager software, also known as Local Run Manager software, is a locally-hosted run management system designed for Illumina benchtop sequencing systems. This easy-to-use solution enables labs to record samples for a run, specify run parameters, monitor status, analyze sequencing data, and view results. Illumina Run Manager software integrates with instrument control software and can be accessed directly on the MiniSeq™ and iSeq™ 100 Systems via a web browser. An off-instrument version that is compatible with the iSeq System, MiniSeq System, MiSeq Series, and NextSeq™ Series is also available. Illumina Run Manager software leverages multiple analysis modules to analyze data, depending on the library preparation kits used for sequencing.

Learn more

Illumina Run Manager software

BaseSpace Sequence Hub for run management

Labs needing a cloud-based run management solution can connect their sequencing systems to BaseSpace Sequence Hub. This cloud-based genomic run management and bioinformatic analysis environment enables efficient run planning, real-time monitoring, and quality control. In addition to run management capabilities, BaseSpace Sequence Hub provides access to a curated and comprehensive menu of point-and-click analysis applications, including DRAGEN secondary analysis apps. Select DRAGEN apps on BaseSpace Sequence Hub are accessible on board Illumina sequencing systems.

BaseSpace Sequence Hub for run management
Cloud-based genomic run management and bioinformatic analysis with tight sequencing instrument integration

Built for: Illumina instrument customers aiming to plan entire sequencing workflows and perform data analysis in an intuitive, graphical environment.

Learn more

BaseSpace Sequence Hub for run management

Secondary analysis solutions

Lab | Run | **Analytics** | Insight

NGS data can be instantly and securely integrated into the Illumina software ecosystem for secondary analysis. The Illumina Connected Software portfolio features accurate, approachable, and comprehensive bioinformatics solutions meeting the needs of a range of bioinformatics expertise and deployment capabilities (Figure 9).

DRAGEN secondary analysis

Illumina DRAGEN secondary analysis provides accurate, comprehensive, and efficient analysis of NGS data. Researchers can leverage the power of DRAGEN secondary analysis through various deployment options, including user-friendly applications on BaseSpace Sequence Hub or on board select Illumina sequencing systems, providing streamlined and intuitive analysis solutions. The software is optimized for NGS data and provides the most accurate secondary analysis, setting new standards for data accuracy. Using Precision FDA Truth Challenge V2 benchmarking data, DRAGEN secondary analysis demonstrated a 99.84% accuracy score in all benchmarked regions for germline variant calls.[1] DRAGEN secondary analysis also provides exceptional accuracy, winning highest accuracy (99.52% accuracy score) in the MHC Region using Precision FDA Truth Challenge V2 benchmarking data.[1] The software analyzes NGS data for a range of sample types and provides a highly accurate, comprehensive, and efficient solution, empowering labs of all sizes and disciplines to do more with their genomic and transcriptomic data. It would take over 30 open-source tools to partially replicate the breadth of functionality within DRAGEN secondary analysis.

DRAGEN secondary analysis
Accurate, comprehensive, and efficient secondary analysis

Built for: Illumina sequencing customers who want to do secondary analysis in house via a variety of access points (on-board, cloud, on-premises) and applications, including genetic disease testing, cancer, and molecular biology research

Comprehensive multi-application analysis

Analyze bulk or single-cell transcriptomes and methylomes on a single platform

Ultra-efficient workflow

Access fast, accurate analysis with hardware acceleration and DRAGEN ORA lossless compression

Award-winning accuracy

Analyze data with the most accurate secondary analysis tool for germline and somatic variant calling

The powerful DRAGEN software uses field-programmable gate array (FPGA) technology to achieve rapid turnaround times.[2] This highly configurable FPGA technology allows for ultra-efficient hardware-accelerated implementations of genomic analysis algorithms, including BCL file conversion, mapping, alignment, sorting, duplicate marking, gene quantification, small variant calling, and gene fusion calling (Figure 9). The flexible programming nature of the FPGAs enables Illumina to develop an extensive suite of DRAGEN application pipelines, with continual improvements and additions to deliver the best possible accuracy, comprehensiveness, and efficiency in analysis.

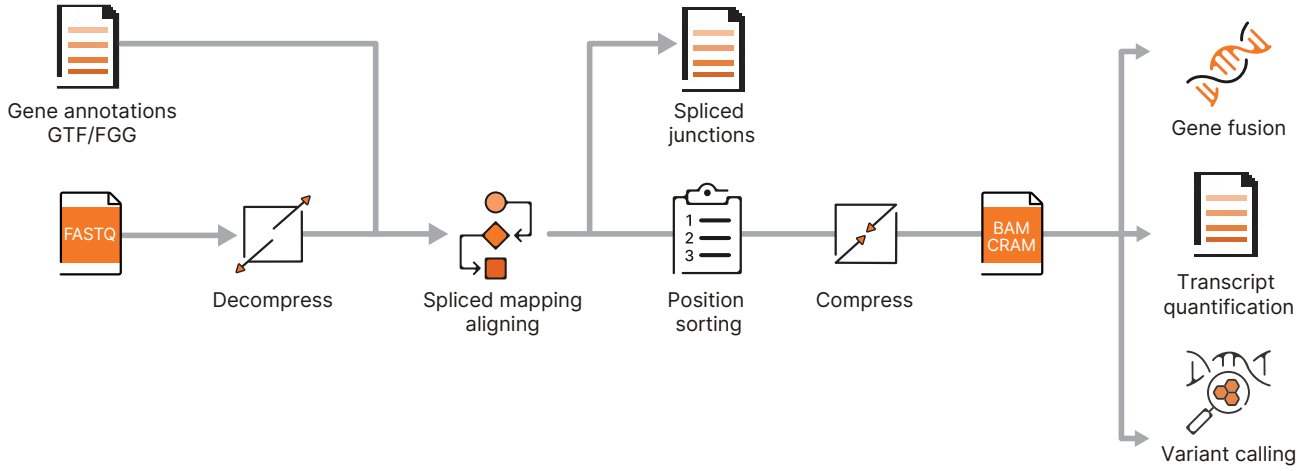


Figure 9: DRAGEN secondary analysis pipelines for RNA-Seq—Each DRAGEN pipeline contains a specific set of steps to support accurate and efficient analysis. DRAGEN secondary analysis provides the flexibility to accept various input files and produce a range of output types, enabling users to customize their experience and produce their desired file format. The DRAGEN RNA pipeline includes an RNA-Seq (splicing-aware) aligner, as well as RNA-specific analysis components for gene expression quantification and gene fusion detection.

Deployment options

DRAGEN secondary analysis pipelines can be accessed through available on-premises, on-instrument, or cloud options, allowing labs to select a solution that best suits their needs (Figure 10). DRAGEN secondary analysis on BaseSpace Sequence Hub enables push-button secondary analysis for labs of all sizes and disciplines. DRAGEN secondary analysis on Illumina Connected Analytics is a comprehensive cloud-based data management and analysis platform that empowers researchers to manage, analyze, and interpret large volumes of multiomic data in a secure, scalable, and flexible environment.

In addition to accessing DRAGEN secondary analysis applications on on the cloud, users can access DRAGEN pipelines on board select Illumina sequencing systems for highly efficient data analysis and an easy-to-use interface.

Deployment option	DRAGEN on-instrument (NextSeq 1000/2000, NovaSeq X Series)	DRAGEN on BaseSpace Sequence Hub	DRAGEN on Illumina Connected Analytics	DRAGEN on-premises server	DRAGEN Multi-cloud Bring Your Own License (AWS, Azure)
Features	Push-button analysis with instrument integration	Push-button solutions with easy-to-use graphical interface	Prepackaged DRAGEN pipelines or tools to incorporate into custom pipelines	Appliance server with command-line interface	Command-line access with license provided by Illumina to run on cloud provider of choice
Configurability	Low	Medium	High	High	High
Frequency of updates	Low	High	High	High	High
Broadness of menu	Low	Medium	High	High	High
Integration with sequencing system	High	Medium	Medium	Low	Low
Push-button analysis	High	High	Medium	Low	Low

Figure 10: DRAGEN pipeline deployment options with features designed to fit the NGS analysis needs of every lab.

Learn more

DRAGEN secondary analysis

Recommended secondary analysis solution for users who are new to NGS

BaseSpace Sequence Hub for DRAGEN secondary analysis and run management

BaseSpace Sequence Hub can be accessed from any internet-connected computer or mobile device and offers a security-first environment that enables researchers to plan and monitor sequencing runs efficiently and leverage powerful DRAGEN secondary analysis pipelines. BaseSpace Sequence Hub provides a cloud-based environment for real-time run monitoring, quality control, and run planning capabilities in addition to access to a curated and comprehensive menu of point-and-click analysis applications, in a single, easy-to-use cloud environment. The intuitive graphic user interface is ideal for individuals with little bioinformatics experience. Push-button out-of-the-box pipelines empowers users of any skill level to perform run management and secondary analysis with confidence and ease.

BaseSpace Sequence Hub for DRAGEN secondary analysis and run management
Efficient cloud-based data analysis applications

Built for: Users who are non-programmers, want to manage runs and analyze experiments using out-of-the-box pipelines and applications

Efficient data integration

Streamlined, intuitive run planning reduces manual touchpoints and opportunities for error

Easy to use

User-friendly, graphical interface simplifies complex tasks and workflows

Intuitive, guided analysis

Curated and comprehensive menu of point-and-click analysis apps accelerate analysis

BaseSpace Sequence Hub features various push-button analysis solutions and version-controlled, ready-to-use workflows. This cloud-based platform 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 (Figure 11), or from a growing network of third-party app providers.

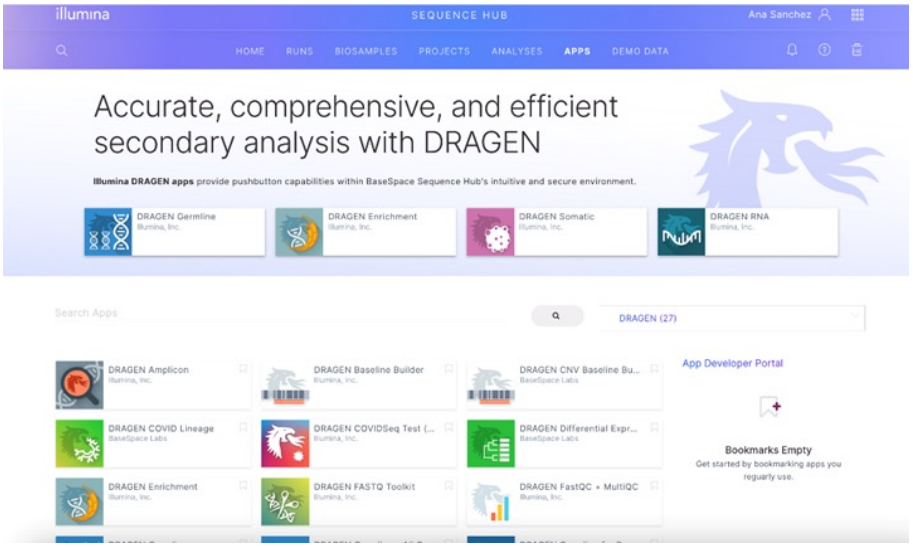


Figure 11: DRAGEN secondary analysis pipelines on BaseSpace Sequence Hub Apps store.

DRAGEN pipelines for gene expression and regulation on BaseSpace Sequence Hub

- DRAGEN RNA pipeline
- DRAGEN Differential Expression app
- DRAGEN RNA Amplicon app

Learn more

BaseSpace Sequence Hub

Illumina Connected Analytics

Illumina Connected Analytics is a secure bioinformatics platform designed to operationalize informatics and drive scientific insights. Leverage the broad flexible applications on Illumina Connected Analytics to implement single-sample workflows, experiment-level discovery, service operations, and more.

With Illumina Connected Analytics, users can build, version, and deploy flexible analytical pipelines while maintaining data privacy, security, and compliance at scale.

Illumina Connected Analytics
Secure and flexible bioinformatics platform to drive scientific insights

Built for: Users looking to aggregate and analyze data at scale, with the flexibility to build and deploy customized analytical pipelines in a cloud-based environment

Build and customize analysis pipelines

Import, build, and edit workflows easily with tools like CWL and Nextflow

Execute production workflows at scale

Interpret your data in a flexible computing environment that includes JupyterLab Notebooks

Explore and share data and results

Organize any data in a secure workspace and share it globally in a compliant manner

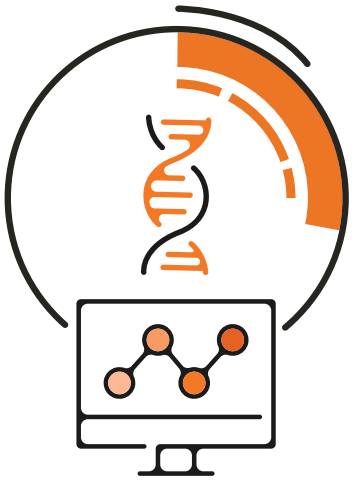
Learn more

Illumina Connected Analytics

Tertiary analysis tools

Lab Run Analytics **Insight**

Tertiary analysis adds critical biological context to the results generated from secondary analysis of NGS data. Depending on the biological question and sequencing method, tertiary analysis can span a broad range of 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 be combined with phenotype information, biological knowledge from functional genomics, or other data sources to increase our 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.



Learn more

Multiomics ebook

Single-cell sequencing ebook

Partek® Flow® software

Partek, an Illumina company, offers easy-to-use bioinformatics software for analyzing and visualizing single cell, gene expression, ChIP-Seq, and other data used for multiomics. Partek Flow software is a robust start-to-finish NGS analysis software designed for researchers of all skill levels, offering an easy-to-use interface (Figure 12), robust statistical algorithms, information-rich visualizations, and cutting-edge genomic tools, enabling users to analyze genomic data confidently. The software enables researchers to produce publication-ready visualizations, collaborate and share customized analysis pipelines, aggregate multiomic and phenotypic data, augment their cohorts to include curated public data sets for well-powered studies, and perform statistical analyses, all on a single platform. Researchers can access out-of-the-box tools for comprehensive data analysis, including genome alignment, differential analysis, QA/QC reports, peak calling, data exploration and quantification, variant annotation, and more.

Partek Flow software
NGS analysis software for rapidly analyzing and visualizing bulk and single-cell multiomics data sets in an easy-to-use interface

Built for: Researchers who want to aggregate and analyze multiomic data in a single place, using publically available statistical algorithms and a range of alignment, QC, data exploration, and quantification tools in a secure, intuitive data environment

Figure 12: Partek Flow software user interface.

Learn more

Partek Flow software

Correlation Engine

Correlation Engine is an interactive omics knowledge base that puts sequencing data in biological context with highly curated public data. Correlation Engine features a massive database, including more than 25,000 studies, with powerful tools to investigate patterns in data and answer queries in real time. Data types support the investigation of both gene expression and gene regulation. In addition to thousands of gene expression studies, Correlation Engine also provides access to curated DNA methylation, DNA copy number, histone modification, ATAC-Seq, and microRNA data. Biogroups representing, for instance, genes with shared regulatory motifs, enable investigation of upstream regulation. With this interactive platform, researchers can automate omics pattern discovery, access a comprehensive suite of applications, and integrate and analyze multiomic data to understand differential gene activity across multiple data types and species (Figure 13). These insights enable data validation and hypothesis testing, so you can get the most out of your gene expression and regulation research.

Correlation Engine
Simplified correlation analysis to determine biological context by leveraging a growing library of curated data sets

Built for: Users who are non-programmers and want to query highly curated public omics data to correlate patterns and contextualize their own data

Figure 13: Correlation Engine—Tools available to interrogate differential expression and investigate patterns in data. Multiple tools enable users to interrogate and interpret their data against public data in a variety of ways, including, for example, to identify knockouts or other gene perturbation models that are positively or negatively correlated to their study signatures in the case of the 'Knockdown atlas'.

Learn more

Correlation Engine

Summary

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 behind gene expression, regulation, and more. Illumina Connected Software simplifies the NGS workflow for researchers just getting started, offering approachable and easy-to-use solutions to streamline the journey from sample to insight. Integrated directly with our sequencing systems, Illumina Connected Software supports genomic researchers from primary to tertiary analysis, optimizes lab and sample management, and accurately calls genetic variations. Providing push-button and intuitive solutions, Illumina Connected Software enables insights for single-sample studies and more. Depending on the experimental question and sequencing method used, an extensive array of informatics tools and pipelines are available.

New to NGS?

Illumina implementation partners can help design and build your informatics solutions to get you up to speed faster. Contact implementation_partners@illumina.com for more information.

Bulk and single-cell approaches for gene expression and regulation research

Researchers can apply either bulk or single-cell NGS approaches to study gene expression and regulation. Each of these methods provides information about gene expression at different levels of resolution (Figure 14). Your research goals will ultimately determine the choice of NGS method used to get the most out of your transcriptomics studies.

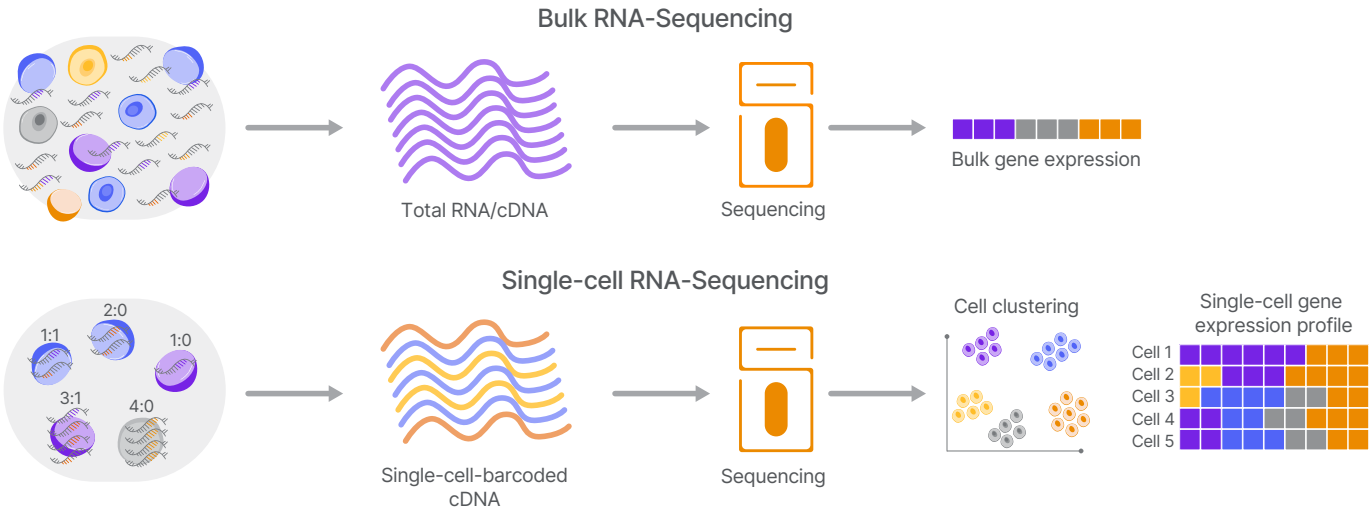


Figure 14: Comparison between bulk and single-cell RNA-Seq approaches—With bulk analysis (top), gene expression is averaged across all the cells included in the sample. However, with scRNA-Seq gene expression (bottom) data are generated for individual cells, enabling deeper insights into the nuanced distinctions between cells within the same sample.

Bulk vs single-cell studies

Conventional bulk RNA-Seq, Methyl-Seq, ChIP-Seq, and ATAC-Seq are the most straightforward approaches to measure the average gene expression or regulation across a pooled sample of cells. These cost-effective methods give researchers comprehensive insights into overall trends in differential gene expression or regulation at scale. However, bulk approaches do not account for the inherent heterogeneity of cells within tissues. Transcripts can be expressed at different levels within a cell population, either due to environmental signals or stochastic changes that occur over time. In addition, low-expressing genes identified in bulk RNA-Seq may be robustly expressed in a rare cell type within that pooled sample of cells.

With the recent advancements in cell isolation and sequencing techniques, it is now possible to study gene expression at the level of individual cells. Single-cell RNA-Seq (scRNA-Seq) and single-cell ATAC-Seq (scATAC-Seq) provide a high-resolution view of cell-to-cell variation, revealing the extensive cellular heterogeneity underlying complex biological systems. Studying these unique characteristics of individual cells delivers valuable insights into cellular processes that might be missed with bulk analyses.

Analyzing bulk gene expression and regulation data

Bioinformatics expertise is no longer a prerequisite for analyzing the large volumes of complex data obtained from NGS-based experiments. Illumina informatics tools and pipelines are highly accessible and can be used by all researchers, even those without prior bioinformatics experience, to analyze data. This section covers the NGS bioinformatics workflow for secondary and tertiary analysis of bulk gene expression and regulation studies.

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 (Figure 15). For differential expression analysis, output files 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.

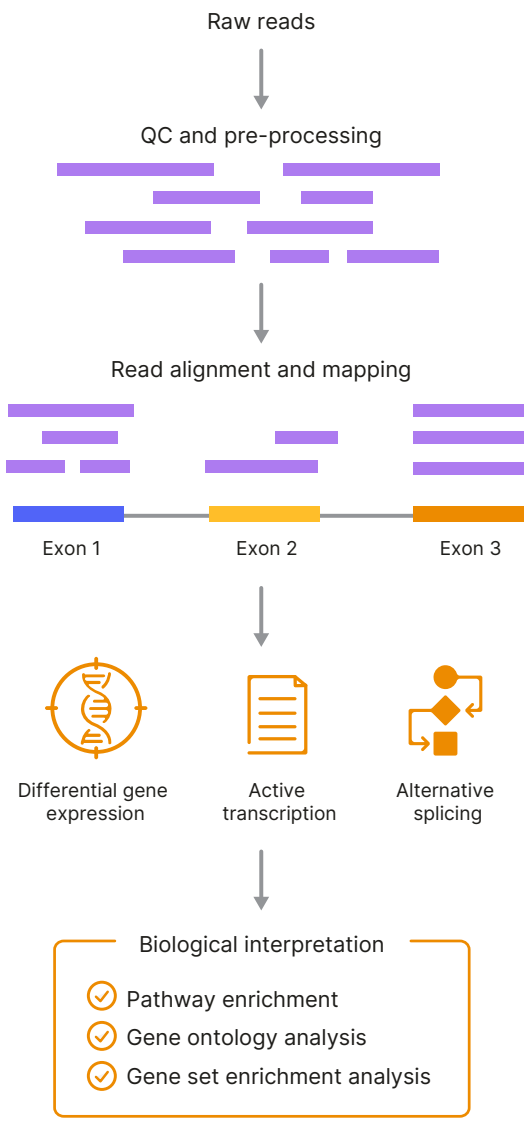


Figure 15: Bulk RNA-Seq analysis steps.

Several analysis tools developed by Illumina, with varying approaches to analyzing gene expression, are available for secondary analysis via DRAGEN secondary analysis, BaseSpace Sequence Hub, and Partek Flow software. Many analysis pipelines begin with alignment (eg, TopHat, STAR, Strelka). Alignment-free approaches (eg, Sailfish, Salmon) are more computationally efficient, but may exhibit lower accuracy with lower-abundance transcripts. 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.

DRAGEN pipelines enable rapid and highly accurate secondary analysis of RNA transcripts

DRAGEN RNA app

The DRAGEN RNA app performs secondary analysis of RNA transcripts and includes a splicing-aware RNA-Seq aligner and RNA-specific analysis components to quantify gene expression and detect gene fusions. This pipeline 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 chimeric alignments to detect potential breakpoints, while adding minimal processing time to the overall pipeline.

Learn more

DRAGEN RNA app

DRAGEN Differential Expression app

The DRAGEN Differential Expression app runs the DESeq2 algorithm on RNA quantification files produced by the DRAGEN RNA app to output genes and transcripts that are differentially expressed between two sample groups. DESeq2 performs differential expression analysis of reference genes on aligned samples to produce gene counts, gene FPKMs (fragments per kilobase of exon per million mapped fragments), principal component analysis, and control vs comparison results.

Learn more

DRAGEN Differential Expression app

DRAGEN RNA Amplicon app

Amplicon sequencing is a highly targeted approach that uses oligonucleotide probes designed to target and capture regions of interest prior to sequencing. The ultradeep sequencing of PCR products, or amplicons, enables efficient variant identification and characterization. The DRAGEN RNA Amplicon app uses the DRAGEN RNA pipeline to set amplicon-specific parameters for fusion calling, including a fusion scoring model trained on RNA amplicon data.

Learn more

DRAGEN RNA Amplicon app

Recommended apps for specific use cases

Gene expression method	Suggested analysis tool
mRNA sequencing	DRAGEN RNA app DRAGEN Differential Expression app
Targeted RNA sequencing	DRAGEN Amplicon app
Small RNA sequencing	Small RNA app
Infectious disease sequencing	DRAGEN Targeted Microbial app COVID-19 software Microbiome Metatranscriptomics app

BaseSpace Sequence Hub apps for user-friendly bulk RNA-Seq analysis

Multiple RNA-Seq tools developed by Illumina are available on BaseSpace Sequence Hub. These apps feature an intuitive user-interface, providing functionality to align reads, quantify gene and transcript abundance, call variants for SNVs and small indels, call gene fusion candidates, and provide QC metrics.

Secondary analysis tools for RNA-Seq in BaseSpace Sequence Hub

Software program	Description
Bowtie 2[3]	Aligns short reads
TopHat 2[4]	Aligns RNA-Seq reads, discovers splice sites
TopHat-Fusion[5]	Discovers gene fusions
Cufflinks[4]	Assembles transcripts
Cuffcompare[4]	Compares assemblies to a reference
Cuffmerge[4]	Combines multiple assemblies
Cuffdiff[4]	Performs differential expression analysis
Strelka[6]	Calls small variants
STAR[7]	Maps RNA-Seq reads, detects splice-junctions and gene fusions
Salmon[8]	Quantifies transcript expression
Sailfish[9]	Quantifies RNA isoforms

Partek tools for bulk RNA-Seq analysis

Partek tools can dig deeper into RNA-Seq data to identify alternatively spliced transcripts between experimental conditions. Partek Flow provides powerful gene expression analysis tools, enabling researchers to design more effective experiments by identifying how many samples they need to measure a specific effect. Users can leverage Partek Flow to look for patterns in gene expression change in time series data, build classification models from publishing data, and utilize the integrated genome browser to visualize isoform expression results (Figure 16).

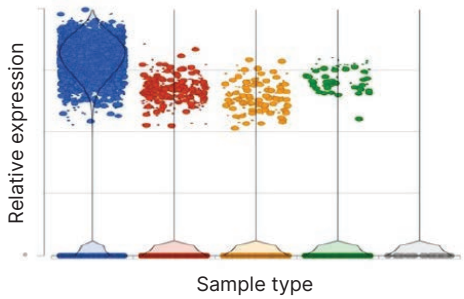
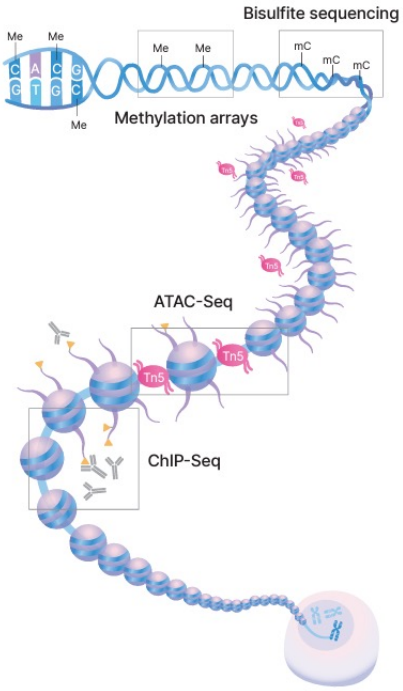


Figure 16: Sample output for differential expression analysis generated using Partek tools—Partek Flow enables users to detect differential gene expression by type using a dot/violin plot.

Secondary analysis tools for RNA-Seq in Partek

Software program	Description
Partek E/M	Tools to quantify transcripts to annotation model or reference
HTSeq	Tools for processing high-throughput sequencing data
GSNAP	Short-read aligner (> 14 bp) capable of handling splicing using either a probabilistic model or database
HISAT2	Alignment program for fast, sensitive mapping NGS reads (DNA and RNA) to a population of genomes
Issac2	Gapped aligner that finds candidate mapping positions

Bioinformatics solutions for bulk gene regulation studies



Gene regulation is the process of how gene expression is turned on or off in physiological or pathological states. 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, and 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.

ChIP-Seq

ChIP-Seq is a powerful, unbiased method for identifying genome-wide DNA binding sites for transcription factors and other proteins. This approach enables genome-wide surveys of gene regulation. The ChIPSeq app, available via BaseSpace Sequence Hub, optimally analyzes ChIP-Seq data. The app uses MACS2 to identify enriched regions pulled down by chromatin immuno-precipitation, and HOMER to discover motifs within these regions. The raw outputs of these tools are presented in a downloadable interactive table. Partek Flow software can be used to analyze ChIP-Seq data integrating MACS2 and MACS3, while enabling the detection of *de novo* motifs and the search for known motifs. Once peaks and TF binding sites are identified using the Partek Flow toolkit, peaks can be annotated to the genome to explore gene section breakdown, including transcription start sites. A list of target genes generated from ChIP-Seq data can be compared with a list of differentially expressed genes generated from RNA-Seq data using a Venn diagram. The Partek Flow chromosome view allows visualization of peaks and differentially expressed genes together.

ATAC-Seq

ATAC-Seq is a powerful epigenetic discovery tool for determining chromatin accessibility. By sequencing regions of open chromatin, ATAC-Seq helps researchers understand how chromatin packaging and other factors affect gene expression. After initial processing and alignment with Bowtie 2, ATAC-Seq data can be analyzed using various commercial and open-source algorithms and software programs, including the ENCODE ATAC-Seq pipeline, which primarily functions 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) regulation without the need for bioinformatics expertise. Partek Flow software enables users to build a start-to-finish ATAC-Seq and other genomic locus analysis pipelines, providing a point-and-click interface for ease of use.

Methylation sequencing

Cytosine methylation is an important process by which gene expression is regulated at the cellular level. Typically, DNA methylation represses transcription and loss of methylation results in gene activation. The following applications can be used to analyze methylation sequencing data.

DRAGEN Methylation pipeline

The DRAGEN Methylation pipeline is designed for ultra-rapid analysis of whole-genome and targeted bisulfite DNA sequence data. The workflow performs alignment and methyl calling and calculates alignment and methylation metrics. This app supports libraries prepared using many commercially available methylation library preparation kits. TET-assisted pyridine borane sequencing (TAPS) type libraries may be analyzed by enabling the TAPS setting.

Learn more

DRAGEN Methylation pipeline

MethylKit app

The MethylKit App is an R package available on BaseSpace Sequence Hub that analyzes sequencing data for differences in methylation between samples. This apps supports targeted data and common epigenomics analysis tasks such as methylation calling, analysis of differential methylation between samples, and categorization of significant methylation regions. The MethylKit app is designed for push-button use, and is accessible to any researcher, regardless of bioinformatics experience.

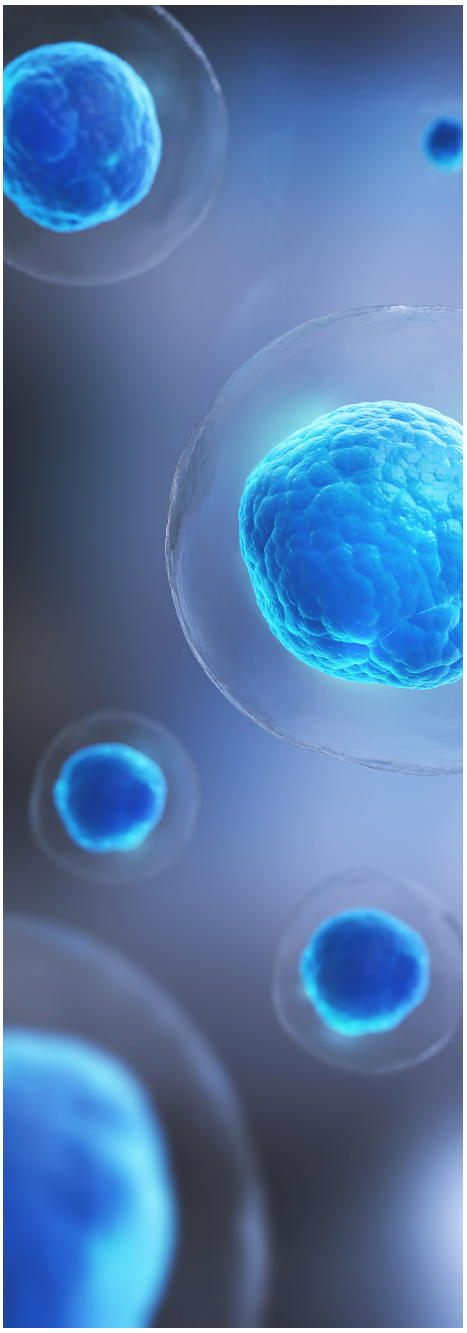
Learn more

MethylKit app on BaseSpace Sequence Hub

Summary

Accessible, user-friendly bioinformatics tools are essential to derive biological insights from bulk gene expression and regulation studies. Illumina data analysis tools support a broad range of bulk gene expression and regulation research applications. When coupled with NGS instruments from Illumina, these easy-to-use data analysis solutions enable researchers to explore gene expression and regulation without the need for bioinformatics expertise.

Analyzing single-cell gene expression and regulation data



Single-cell sequencing data analysis workflows, such as those for scRNA-Seq or scATAC-Seq, are similar to their complementary bulk analysis protocols. However, there are several key considerations that are specific to single-cell sequencing study design, resulting in distinct data features that require specialized bioinformatics tools for optimal analysis.[10,11]

Challenges with single-cell sequencing data analysis

Analysis of scRNA-Seq data presents several challenges not applicable to bulk sequencing data analysis.[12] With advances in technology enabling high-throughput single-cell isolation, the increasing numbers of cells analyzed translates into significantly more data points, requiring scalable analysis methods. Increased variability as compared to bulk RNA-Seq is unique to scRNA-Seq data, which is reflective of this increase in biological complexity. Further complicating scRNA-Seq data analysis are observed zeroes or "dropouts," which are genes for which no unique molecular indentifiers (UMIs) or sequencing reads map to a particular cell. This phenomenon can occur either from technical issues, where the gene is expressed but is not detected, or the gene truly is not expressed by that particular cell. Another added layer of complexity is the temporal nature of gene expression, in which 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.[13]

Key features of single-cell data analysis

The analysis pipeline for single-cell sequencing data involves aligning reads, generating feature-barcode matrices, and performing tertiary analysis. Primary analysis for single-cell sequencing data is similar to that for bulk sequencing, consisting primarily of file conversion from the BCL to FASTQ format. The following steps in secondary and tertiary analysis are distinct from bulk RNA-Seq if data from single-cells is being analyzed.

Secondary analysis

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 17). The powerful DRAGEN Single-Cell and DRAGEN Single-Cell ATAC pipelines include demultiplexing capabilities for optimized downstream analyses.

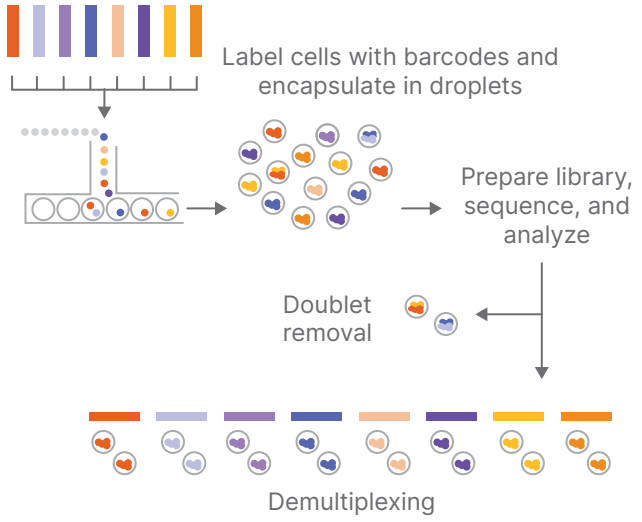


Figure 17: Demultiplexing in single-cell sequencing—Cell barcodes added during isolation are used to parse sequencing reads to individual cells and remove cell doublets during secondary analysis.

QC metrics

Before downstream analysis, several QC metrics can be used 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. 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.

Learn more

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

Tertiary analysis

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. 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.[14] Dimensionality reduction enables clustering of data points and is an important step in the analysis of scRNA-Seq, scATAC-Seq, and single-cell protein profiling.

Principal component analysis (PCA)

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 18). PCA assumes data are approximately normally distributed, which may not always apply to single-cell sequencing data.

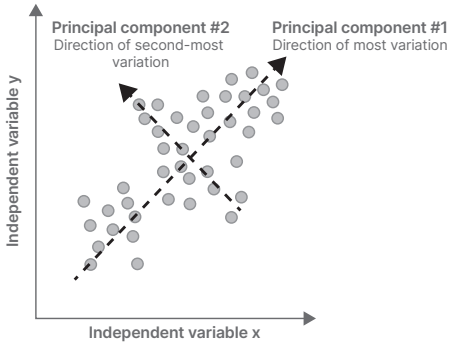


Figure 18: Clustering with PCA—PCA projects high dimensional data to a lower number of dimensions for maximal variance capture but is restricted to linear dimensions and assumes normally distributed data.

t-distributed stochastic neighborhood embedding (t-SNE)

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 19).

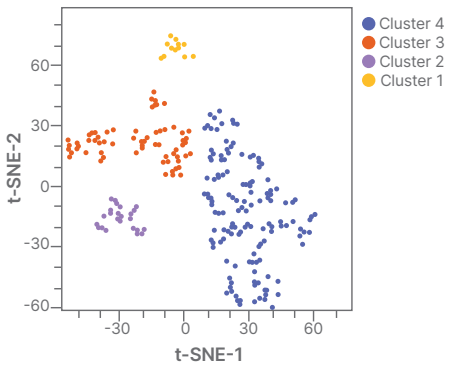


Figure 19: Clustering with t-SNE—t-SNE reduces dimensionality, plotting data points on a two- or three-dimensional plot. Importantly, while clusters denote similarity between data points, the distance between clusters does not indicate similarity, that is, clusters 1 and 3 are not necessarily more similar than clusters 1 and 2.

Uniform manifold approximation and projection (UMAP)

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.

Bioinformatics tools for single-cell gene expression studies

Powerful analysis tools developed by Illumina, with varying analysis approaches to analyzing gene expression in single-cells, are available for secondary analysis via DRAGEN secondary analysis and BaseSpace Sequence Hub. The Illumina single-cell RNA analysis workflow begins with file conversion from BCL to FASTQ, followed by secondary analysis using the DRAGEN Single-Cell app. Interactive tertiary analysis tools available through the Illumina Connected Analytics platform enable researchers to process and visualize their results, placing single-cell data into a biological context. Partek Flow software also enables users to analyze single-cell data from start to finish, combining powerful statistics with information-rich, interactive visualizations in a point-and-click interface.

DRAGEN Single-Cell RNA app

The DRAGEN Single-Cell RNA app can process multiplexed scRNA-Seq data sets from reads to a cell-by-gene UMI count gene expression matrix (Figure 20). The pipeline is compatible with library designs that have one read in a fragment match to a transcript and the other containing a cell-barcode and UMI. The functionality and options related to alignment and gene annotation are identical to the RNA-Seq pipeline and can be accessed via prepackaged pipelines on BaseSpace Sequence Hub and Illumina Connected Analytics, or on an on-premises server. The following functions are included:

- RNA-Seq (splice-aware) alignment and matching to annotated genes for the transcript reads
- Cell-barcode and UMI error correction for the barcode read
- UMI counting per cell and gene to measure gene expression
- Cell hashing and feature counting by read 2 UMI
- Sparse gene expression matrix output and scRNA QC metrics

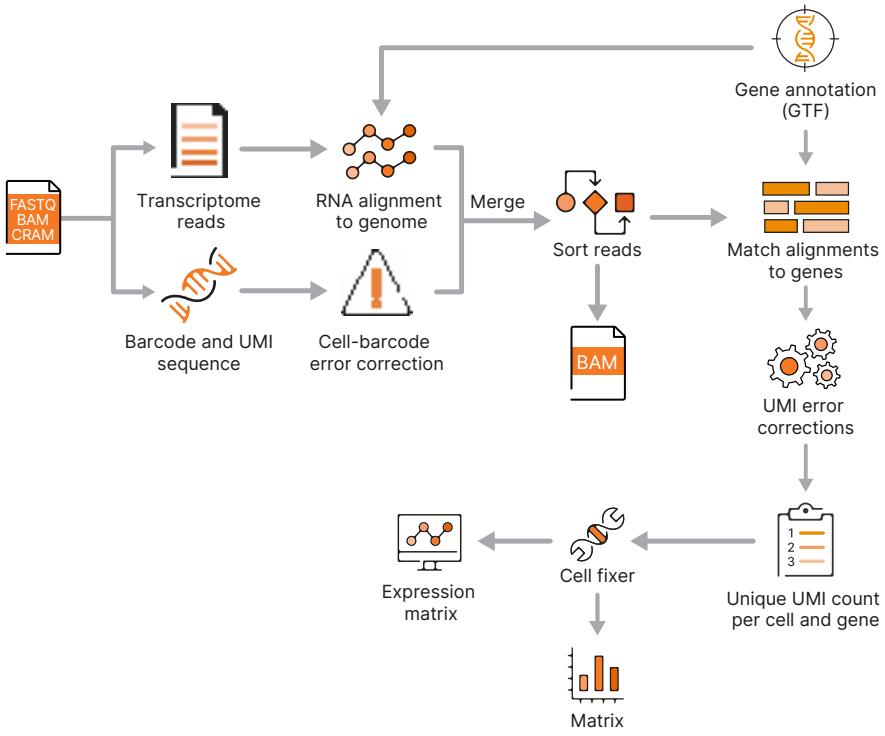


Figure 20: DRAGEN Single-Cell pipeline workflow.

Learn more

DRAGEN Single-Cell RNA app

Partek Flow single-cell RNA-Seq analysis

Partek Flow provides several tools to power single-cell resolution profiling of gene expression and chromatin accessibility (Figure 21). Tools can be run in combination, enabling simultaneous high-resolution profiling of the transcriptome and epigenome.
The following features are included within Partek Flow software:

- Automatic or manual cell classification
- Cell selection and filtration on t-SNE scatter plots
- Graph-based clustering
- Powerful visualizations, including PCA, UMAP, and heatmaps
- Gene ontology (GO) and pathway enrichment analysis

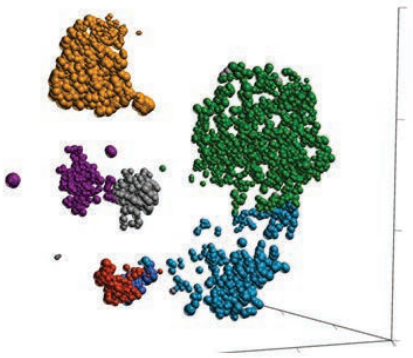


Figure 21: Sample output for single-cell classification using Partek—Partek Flow can classify cells by traditional methods or automatically. Each cluster shown here corresponds to a specific cell population within the sample.

Powerful tools and features in an easy-to-use application:

- Automatic or manual cell classification
- Interactive visualization
- Powerful industry-standard statistics
- Easy data import and export
- Robust QA/QC capabilities
- Multi-sample analysis
- Flexible differential analysis
- Complete and unbiased biological interpretation
- Biomarker discovery
- Comparison of cell populations between experimental groups

Learn more

Partek Flow software for single cell analysis

Tertiary analysis on Illumina Connected Analytics

The Illumina Connected Analytics Bench module features customizable tools for tertiary analysis of data from DRAGEN secondary analysis pipelines. This module computes data using common scripting languages, including R and Python. Users can interpret data in a flexible computing environment that supports JupyterLab Notebooks and R Shiny apps. For simplified analysis, researchers can access the Single-Cell app within the Illumina Connected Analytics Bench module. The R Shiny application, developed by Illumina bioinformaticians, performs clustering, cell-type identification, differential expression analysis, and scRNA-Seq data visualization. The step-by-step guide simplifies the process of analyzing output files from the DRAGEN Single-Cell RNA app and creating customizable reports with publication-quality images.

Learn more

Illumina Connected Analytics—Interactive analysis for single cell RNA workflow video

Bioinformatics solutions for single-cell gene regulation studies

DRAGEN Single-Cell ATAC app

Analyzing scATAC-Seq experiments can be challenging because the data sets tend to be large, sparse, and binary. The DRAGEN Single-Cell ATAC app enables single-cell resolution profiling of chromatin accessibility. This powerful tool can be run with the DRAGEN Single-Cell RNA app as part of the single-cell multiomics pipeline, enabling simultaneous high-resolution profiling of the transcriptome and epigenome. The following features are included:

- ATAC-Seq alignment with barcode error correction
- Chromatin accessibility peak calling
- Fragment counting per cell and peak to measure chromatin accessibility
- Sparse read count matrix output and scATAC-Seq QC metrics

Learn more

DRAGEN Single-Cell ATAC app

Partek Flow single-cell ATAC-Seq analysis

Partek Flow provides powerful gene expression analysis tools, enabling researchers to analyze and visualize multiple samples together or independently and discover biological variations with combined and split-by-sample analysis. The following features are included within Partek Flow:

- Automatic or manual cell classification
- TF-IDF normalization and SVD
- Grah-based clustering, UMAP
- Promoter sum matrix output and scATAC-Seq QC metrics

Learn more

Partek Flow software single cell resource guide

Summary

Single-cell sequencing data present unique challenges for analysis, requiring specialized software to maximize insights into cellular heterogeneity at high resolution. The ability to derive meaningful insights from these methods is key to understanding the complexity of cellular and molecular biology. Innovative and user-friendly bioinformatics tools developed by Illumina simplify analysis, visualization, and interpretation for scRNA-Seq, scATAC-Seq, and protein profiling studies. These high-performance solutions empower researchers to get the most out of their gene expression and regulation data.

Unlock insights with Illumina Connected Software

Gene expression and regulation studies provide a snapshot of actively expressed genes and transcripts under various conditions. Over the past decade, the scope of transcriptomics studies has evolved from interrogating a few genes at a time to assessing genome-wide expression levels in a single experiment. Accurate interpretation of the massive amounts of variant data generated during these large-scale experiments is a significant challenge for laboratories studying gene expression and regulation. Illumina Connected Software simplifies data analysis, maximizes efficiency, and empowers researchers to transform data into insights. Powerful tools for configuration, seamless instrument integration, and intuitive user interfaces mean that deep bioinformatics expertise is no longer a prerequisite for analyzing the large volumes of complex data obtained from NGS-based experiments.

Illumina Connected Software simplifies the process of gaining insights from gene expression and regulation data, providing highly accurate and intuitive solutions for every step of the NGS workflow, from sample management to insights. The Illumina software portfolio features secure, versatile solutions with award-winning accuracy, direct integration with Illumina sequencing systems, and push-button workflows with intuitive user interfaces. Illumina Connected Software supports a wide range of applications, including differential gene expression, transcriptome profiling, and more, best enabling biologists to maximize the discovery power of their gene expression studies.

Learn more

Illumina Connected Software

Abbreviations

API: application program interface

ATAC-Seq: assay for transposase-accessible chromatin using sequencing

BAM: binary alignment map

BCL: binary base call

BED: browser extensible data

cDNA: complementary DNA

ChIP-Seq: chromatin immunoprecipitation assay with sequencing

CLI: command line interface

CNV: copy number variant

DNA: deoxyribonucleic acid

DRAGEN: Dynamic Read Analysis for GENomics

FPGA: field-programmable gate array

GUI: graphical user interface

LIMS: laboratory information management system

MACS: model-based analysis of ChIP-Seq

mRNA: messenger RNA

NGS: next-generation sequencing

ORA: original read compression

PCA: principal component analysis

QC: quality control

RNA: ribonucleic acid

RNA-Seq: RNA sequencing

RTA: real-time analysis

SAM: sequence alignment map

SBS: sequencing by synthesis

scATAC-Seq: single-cell ATAC-Seq

scRNA-Seq: single-cell RNA sequencing

SNP: single nucleotide polymorphism

SNV: single nucleotide variant

SV: structural variant

TAPS: TET-assisted pyridine borane sequencing

t-SNE: t-distributed stochastic neighborhood embedding

UMAP: uniform manifold approximation and projection

UMI: unique molecular identifier

VCF: variant call format

References

1. PrecisionFDA. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions - precisionFDA Challenge. https://precision.fda.gov/challenges/10. Accessed June 7, 2023.
2. Miller NA, Farrow EG, Gibson M, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med.* 2015;7(1):100. doi:10.1186/s13073-015-0221-8
3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359. doi:10.1038/nmeth.1923
4. 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. doi:10.1038/nprot.2012.016
5. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 2011;12(8):R72. doi:10.1186/gb-2011-12-8-r72
6. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma Oxf Engl.* 2012;28(14):1811-1817. doi:10.1093/bioinformatics/bts271
7. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl.* 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
8. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods.* 2017;14(4):417-419. doi:10.1038/nmeth.4197
9. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol.* 2014;32(5):462-464. doi:10.1038/nbt.2862
10. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 2016;17(1):63. doi:10.1186/s13059-016-0927-y
11. 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. doi:10.1186/s13073-017-0467-4
12. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biol.* 2020;21(1):31. doi:10.1186/s13059-020-1926-6
13. Gao M, Ling M, Tang X, et al. Comparison of high-throughput single-cell RNA sequencing data processing pipelines. *Brief Bioinform.* 2021;22(3):bbaa116. doi:10.1093/bib/bbaa116
14. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol.* 2019;37(1):38-44. doi:10.1038/nbt.4314

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel

techsupport@illumina.com | www.illumina.com

M-GL-00060 v1.0