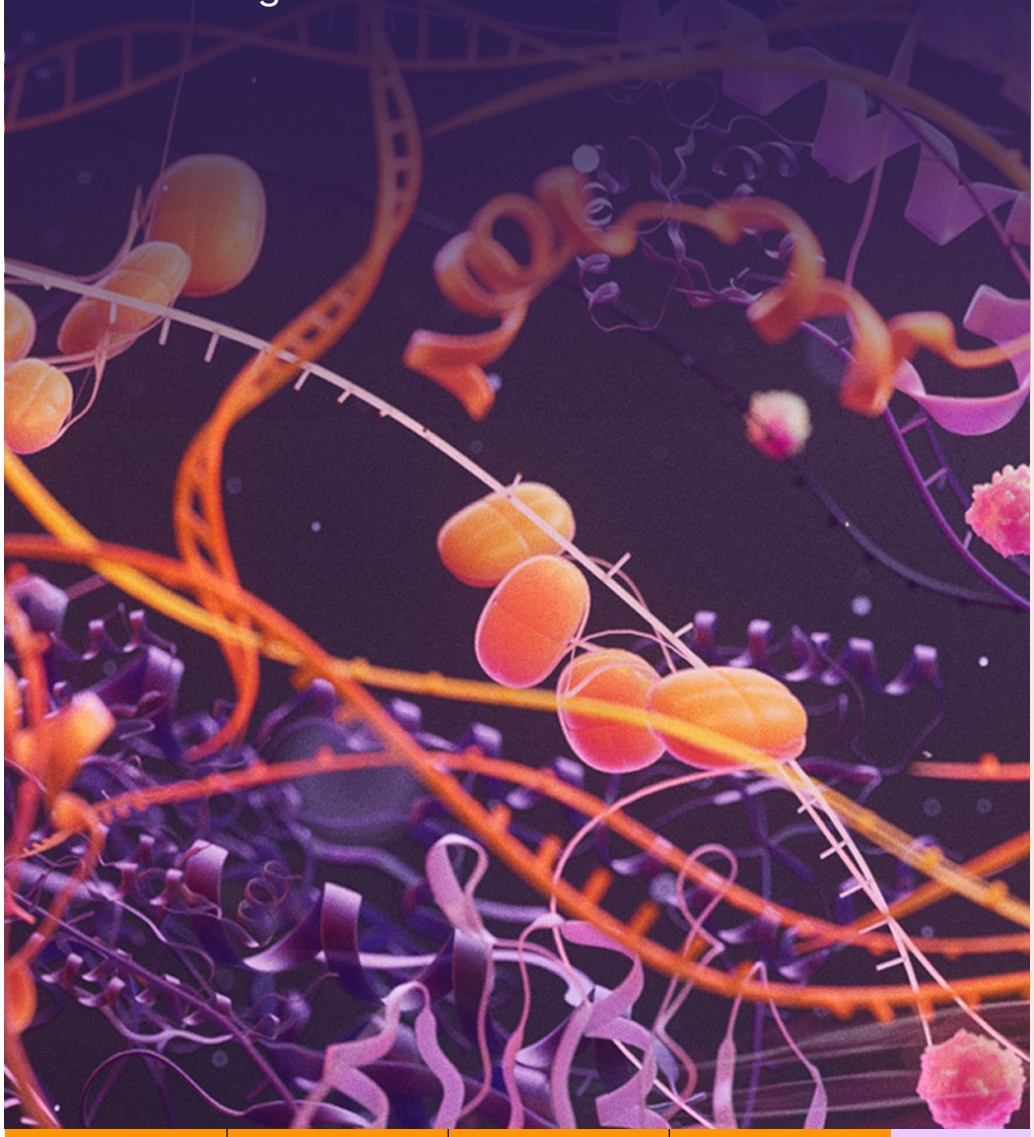
More with multiomics: An introduction to multiomic data analysis

Explore the multiomic data analysis pathway, as illustrated by genomic and transcriptomic data integration



GENOMICS | EPIGENOMICS | TRANSCRIPTOMICS | PROTEOMICS

illumina®

M-AMR-01553 v2.0

Table of contents

Introduction

Next-generation sequencing (NGS) technologies have revolutionized our understanding of life's complexity by enabling deep sequencing of the genome, epigenome, transcriptome, proteome, and more.[1] To gain a more comprehensive view of biological systems, researchers are increasingly adopting multiomic approaches that integrate data across multiple molecular layers (omes) (Figure 1).[2-4] These approaches continue to expand with emerging technologies such as spatial transcriptomics, which combines gene expression measurements with spatial information to provide additional biological context. Published literature and market analyses indicate increasing adoption of multiomic approaches as sequencing technologies become more accessible and cost-effective.[5] The Illumina MiSeq™ i100 System, NextSeq™ 1000 and NextSeq 2000 Systems, NovaSeq™ 6000 System, and NovaSeq X Series offer the application flexibility to enable cost-effective multiomic approaches.

After generating sequencing data from multiple omes, scientists require analysis workflows that provide accurate and reproducible results. Integrating data across multiple omes generates powerful insights into life's most fundamental processes, creating a comprehensive picture of biological phenomena. This eBook explores the integration of genomic and transcriptomic sequencing data as an example of a multiomic data analysis workflow.

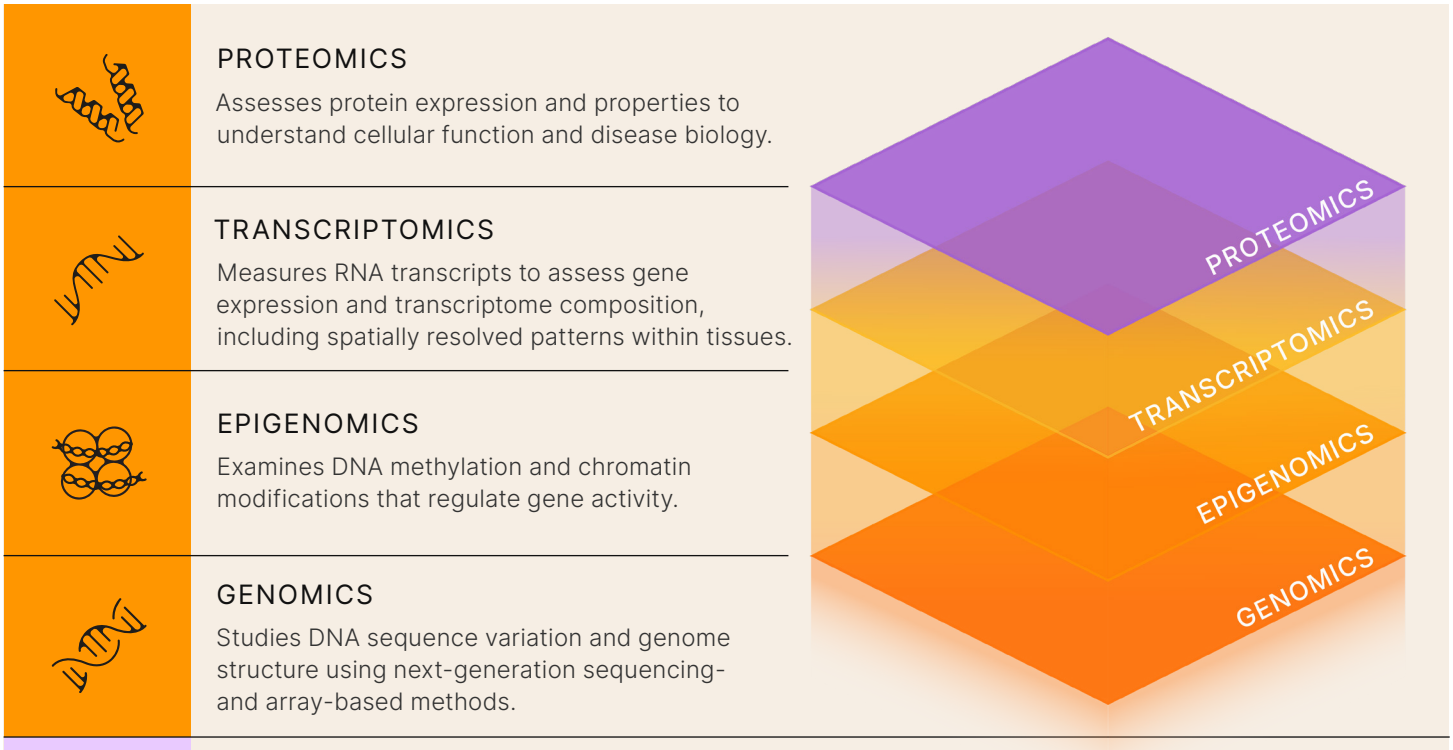


Figure 1: Multiomics connects information across multiple layers to create a comprehensive view of complex systems.

From sequencing to biological insight

Although multiomic study designs vary widely, most genomics and transcriptomics workflows follow a common analytical framework. Even within this general framework, data analysis has traditionally been complex and labor-intensive. Technological advancements have streamlined these workflows, reducing the burden of maintenance and routine pipeline management and enabling greater focus on scientific interpretation and innovation.[6-7] Illumina cloud-based and on-instrument analysis applications are designed to support standardized workflows and facilitate collaboration across research teams.

A typical data analysis workflow begins with sequencing data in FASTQ format (Figure 2). Reads are assessed, filtered, and aligned to a reference genome to generate Binary Alignment Map (BAM) files. Downstream analysis then diverges by data type: genomic BAM files are processed through variant calling to generate Variant Call Format (VCF) files while transcriptomic BAM files are quantified and normalized to measure gene expression. Finally, genomic and transcriptomic results can be integrated to support visualization, interpretation, and biological discovery. This eBook explores each stage of the data analysis workflow in greater detail.

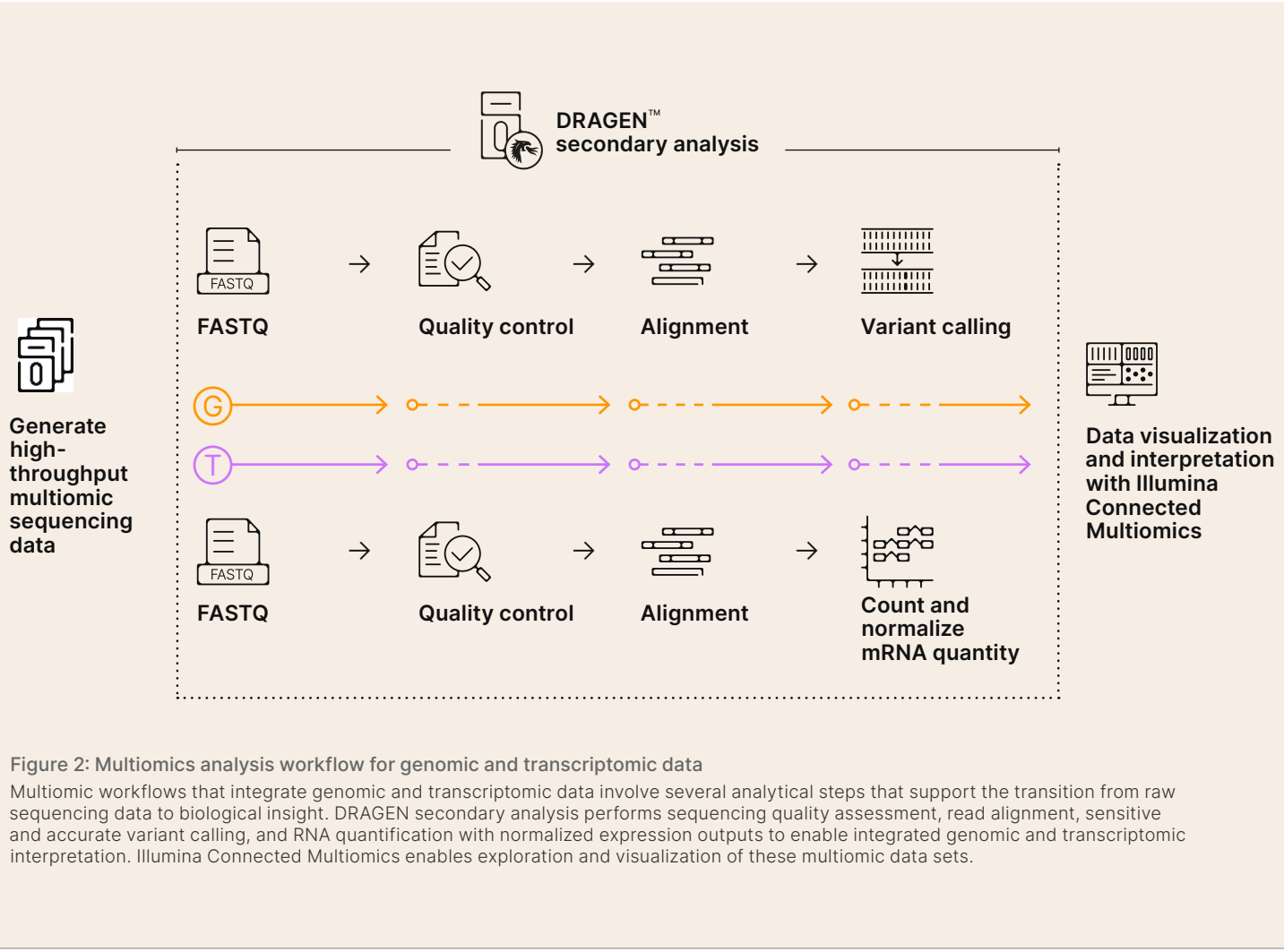


Figure 2: Multiomics analysis workflow for genomic and transcriptomic data

Multiomic workflows that integrate genomic and transcriptomic data involve several analytical steps that support the transition from raw sequencing data to biological insight. DRAGEN secondary analysis performs sequencing quality assessment, read alignment, sensitive and accurate variant calling, and RNA quantification with normalized expression outputs to enable integrated genomic and transcriptomic interpretation. Illumina Connected Multiomics enables exploration and visualization of these multiomic data sets.

The multiomic data analysis workflow

From FASTQ to BAM with DRAGEN software

Quality assessment

Following a run, sequencing systems generate a binary base call (BCL) file. BCL files are commonly converted by core facilities and service providers into FASTQ files, which contain a list of the nucleotide sequences alongside metrics, such as quality scores. This text-based FASTQ file is the standard starting point for most research teams.

Assessing the quality of reads within the FASTQ file is a crucial first step in data analysis for any NGS experiment. Quality assessment includes evaluating average read length, average base quality, and guanine/cytosine content. Assessing these metrics helps identify low-quality reads that may negatively affect downstream analyses. Illumina DRAGEN secondary analysis streamlines quality assessment thanks to push-button options that generate easy-to-understand graphs (Figure 3 and Figure 4).

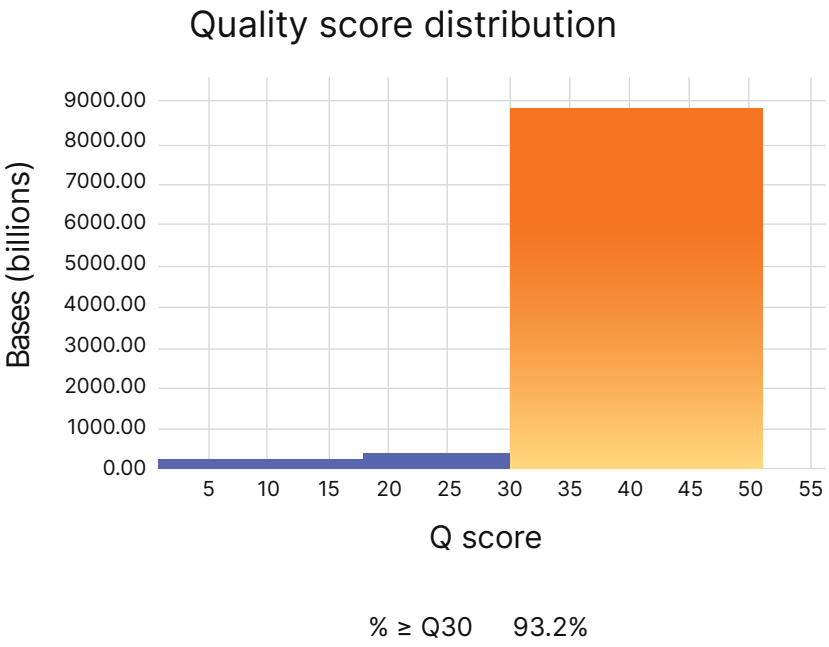


Figure 3: A sample graph illustrating the quality of sequencing data.

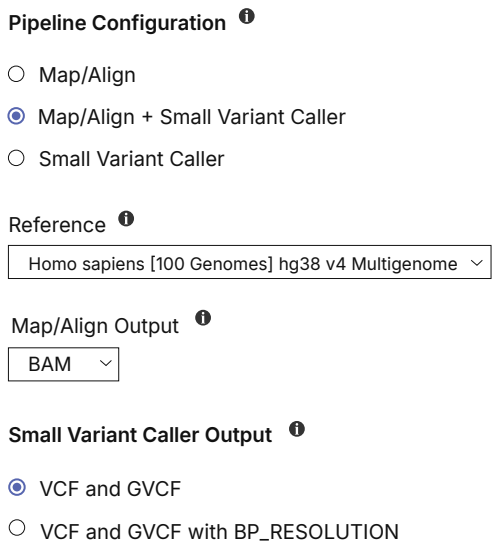
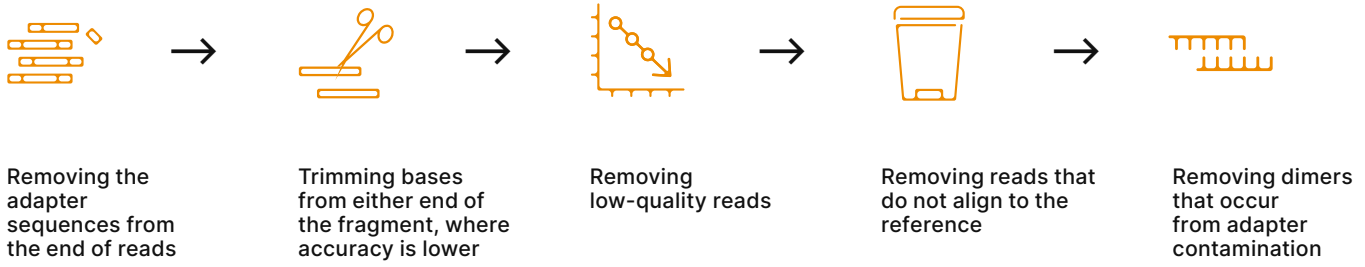


Figure 4: Example of the easy-to-use, push-button interface for configuring DRAGEN secondary analysis. Available interface and configuration options may vary by software version.

Read filtering

After assessing the quality of sequencing reads in the FASTQ file, irrelevant and low-quality sequences are filtered out of the analysis. This includes:



Genome alignment

After low-quality reads have been filtered out, the sequenced fragments can be aligned to a reference genome. Alignment accuracy depends on both sequencing quality and the reference genome used, particularly in regions with high genetic diversity or complex structural variation. These aligned reads are stored in a BAM file for downstream genomic and transcriptomic analyses.

It is important to note that many publicly available reference genomes do not fully capture ancestral and global population variation adequately due to limitations in the sample groups.[8] Furthermore, while many programs use a single reference for genome mapping and alignment (also known as a linear reference), this approach can struggle to resolve highly polymorphic regions of the genome or structural variants such as insertions, inversions, and translocations.[9]

Better alignment with multigenome mapping

One way to achieve a more robust analysis with improved resolution in complex and highly polymorphic regions is to use multigenome mapping.[9] Unlike traditional methods that use one reference genome to align reads, multigenome mapping can use many genomes (pangenome) as a reference source.

Multigenome mapping is gaining adoption across genomics applications because it can:

- Better represent genetic diversity and reduce ancestry-related reference bias[10]
- Improve detection of certain single nucleotide variants (SNVs) compared with linear reference approaches[9]

Multigenome mapping is increasingly represented in the scientific literature and is readily available in DRAGEN secondary analysis.[11]

Independent analysis of genomic and transcriptomic data

Analyzing the genomic BAM file: variant calling

After the BAM file is generated, DNA sequencing data is converted to a VCF file through a process known as variant calling.[12] Variant calling allows researchers to identify mutations, deletions, and insertions by comparing their sequence to one or more reference genomes. Variant detection is influenced by sequencing depth and alignment quality (Figure 5).

Variant calling can identify both germline and somatic mutations. Germline mutations are present in all cells of an individual. As a result, germline variant calls typically show allele frequencies near 0, 0.5, or 1. Somatic mutations, in contrast, arise in nongermline cells and may occur at lower allele frequencies due to biological and sample heterogeneity. Variant detection thresholds may therefore be adjusted to support detection of lower-frequency variants. DRAGEN secondary analysis incorporates machine learning–based approaches designed to support improved variant calling accuracy.

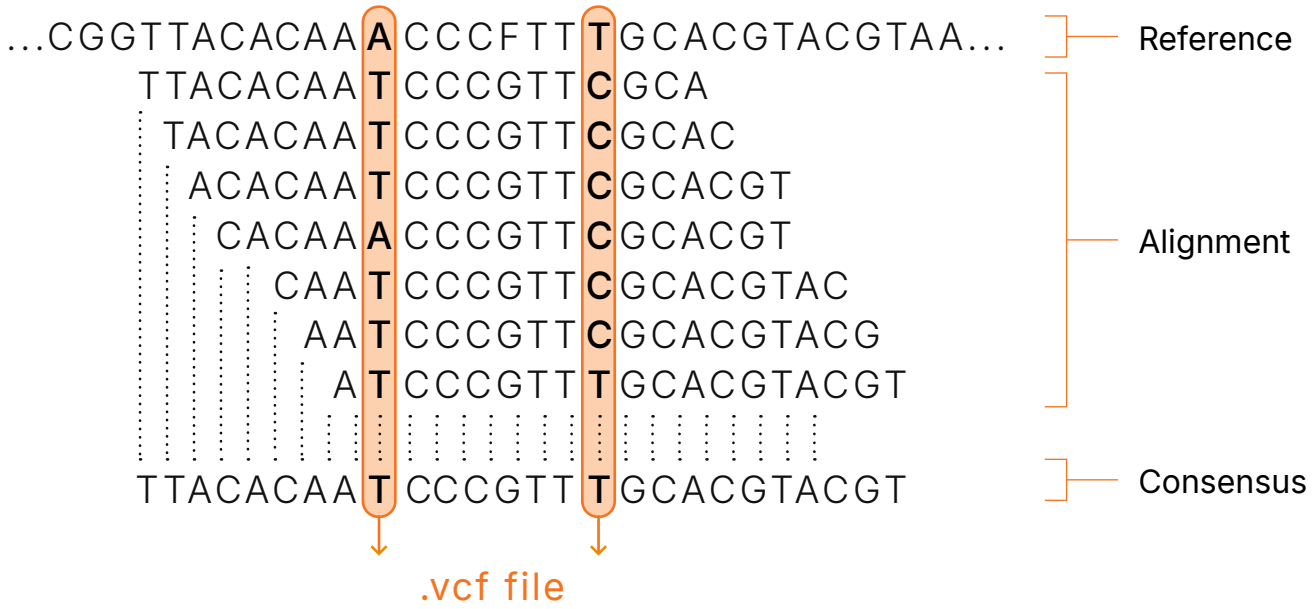


Figure 5: An example of the variant calling process, resulting in a consensus sequence.

Maximizing accuracy: make every base count

Quality assessment of the FASTQ file helps ensure high-quality input into the data analysis pipeline. Maintaining that quality throughout subsequent analysis is equally important. The program used to process sequencing data can have a significant impact on downstream accuracy. If the analysis program is not reliable, variant calls and expression levels may be compromised by false negatives and false positives, especially in regions of the genome that are difficult to sequence.

In the benchmark evaluated, DRAGEN secondary analysis (v4.5) demonstrated fewer variant-calling errors than the comparator workflow tested (Figure 6).[13]

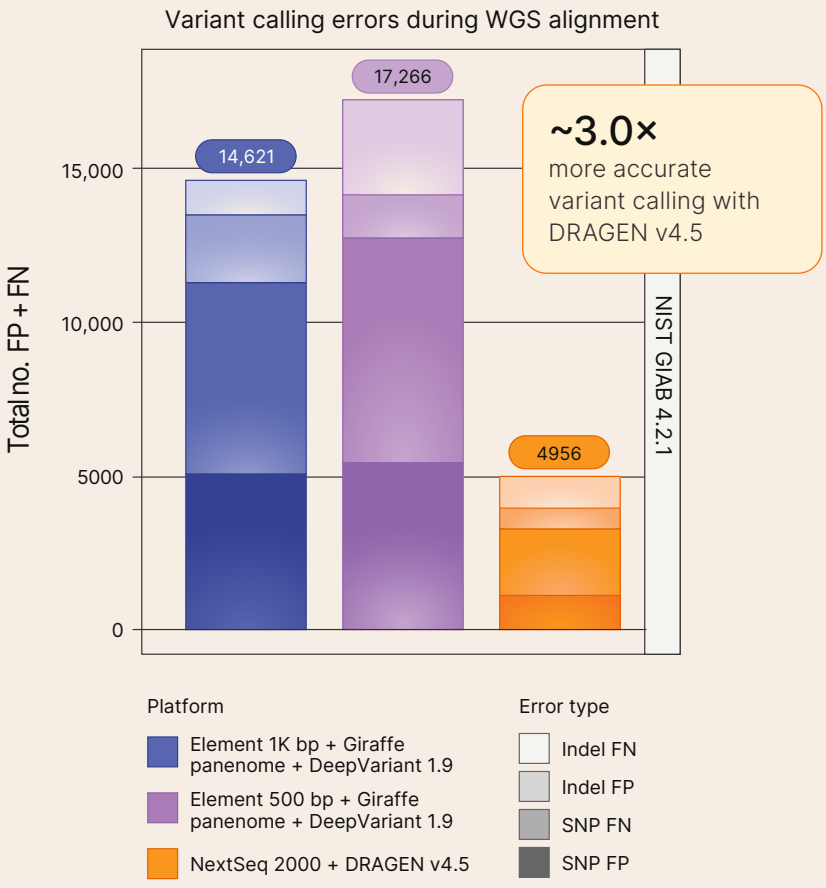


Figure 6: DRAGEN secondary analysis (v4.5) enables approximately 3.0 times more accurate variant calling on the NextSeq 2000 Sequencing System with XLEAP-SBS™ chemistry compared with Element AVITI + DeepVariant based on the National Institute of Standards and Technology (NIST) Genome in a Bottle (GIAB) v4.2.1 benchmark.[13]

Results reflect performance in the NIST GIAB v4.2.1 benchmark under the conditions evaluated and may not predict performance in all workflows, specimens, or applications.

FN, false negative; FP, false positive; indel, insertion/deletion; SNP, single nucleotide polymorphism, WGS, whole-genome sequencing.

Variant refinement and annotation

Following variant calling, variants are filtered to reduce technical artifacts and background noise. The remaining variants are annotated with biological information, including gene and transcript associations, predicted functional effects, population frequencies, and links to curated variant databases. These annotations help researchers organize large variant data sets and prioritize variants that may be relevant to the biological question under investigation.

By combining quality filtering with biological annotation, variant refinement facilitates interpretation of how specific genetic alterations may be associated with observed biological characteristics. Structured variant data can then be compared across samples and integrated with other molecular data sets in downstream multiomic analyses.

Analyzing the transcriptomic BAM File: quantification

After RNA sequencing data have been aligned in a BAM file, the number of reads must be quantified.[14] When sequencing the transcriptome, higher read counts for a given transcript generally corresponds to higher gene expression levels. However, after quantifying the total number of reads, it's important to normalize these counts to account for factors such as gene length, RNA composition, and differences in read depth between samples.

After transcripts have been quantified and the expression levels have been normalized, several statistical analysis methods can be used to discover important patterns in gene expression.

Differential gene expression analysis

One of the most common downstream approaches is the comparison of gene expression patterns across conditions to identify statistically significant differences. Transcriptomic data undergoes processing steps that enable accurate comparison of gene expression across samples and experimental conditions. Expression measurements are filtered and normalized to reduce technical variation and ensure that differences in expression reflect underlying biological signals rather than sequencing or sample-preparation effects. These preprocessing steps establish a consistent foundation for downstream analyses.

Once expression data have been prepared, researchers can examine global transcriptional patterns and identify genes with significantly altered expression levels. Differential expression analysis highlights genes associated with biological responses and states, providing insight into the pathways and processes that drive these transcriptional changes.

Together, these analyses provide a quantitative view of transcriptional activity and support the biological interpretation of transcriptomic data.

Integrated analysis of multiomic data

Multiomic insights emerge when genomic and transcriptomic data sets are explored together to examine how genetic variation influences gene expression. By linking genomic variants with corresponding expression changes, researchers can investigate relationships between genotype and transcriptional activity and assess their potential biological significance. Combining these complementary data types provides a more comprehensive view of molecular function than either data set alone.

Exploring and interpreting data

Illumina Connected Multiomics provides a shared environment for research teams to analyze, visualize, and integrate multiomic data across modalities.[15] Rigorous statistical methods and AI-enabled analytical tools support interpretation of complex multiomic datasets (Figure 7). Researchers can evaluate gene expression patterns, variant associations, pathway activity, phenotype links, and other cross-omic relationships.

The Illumina software and informatics portfolio supports analysis, interpretation, and data aggregation across a range of research applications. Correlation Engine is integrated within Illumina Connected Multiomics, providing biological context for findings as part of the tertiary analysis experience. Researchers can compare gene-level results from transcriptomic analysis against curated public data sets to identify concordant signatures, pathways, regulatory relationships, drug associations, and phenotype links. Correlation Engine mines more than 27,000 scientific studies to deliver data-driven insights for genes, experiments, drugs, and phenotypes, helping place individual experimental results within the broader body of scientific evidence.

Together, integrated analysis and knowledge-driven interpretation transform genomic and transcriptomic data sets into a more comprehensive understanding of complex biological systems.

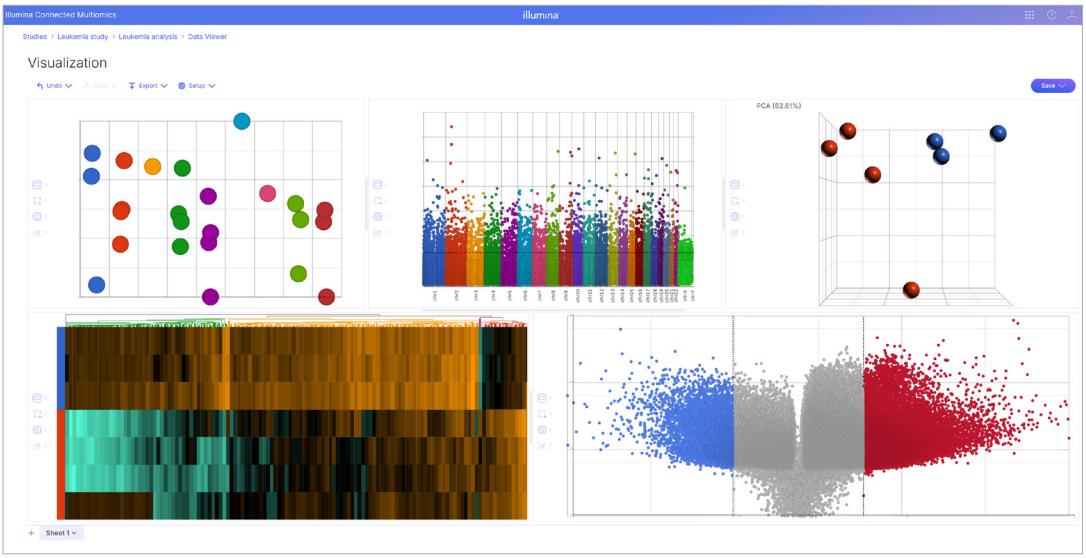


Figure 7: Example visual outputs from Illumina Connected Multiomics include dot plots, Manhattan plots, principal component analysis (PCA) plots, heatmaps, and volcano plots to support evaluation of expression patterns, association signals, sample structure, and differential features across multiomic datasets.

"**Today, multiomics is more accessible.** That's really what it is—it's accessible. The workflows are streamlined in a way that wasn't possible even five years ago. When the workflows are streamlined and the technologies are consolidated so they work together seamlessly, then you can ask the questions, and you're not as intimidated by it because it's doable."

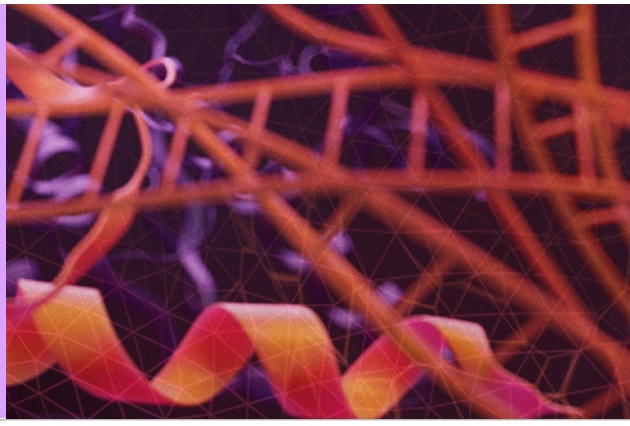
Bodour Salhia, PhD
Professor of Cancer Biology
Royce and Mary Trotter Chair in Cancer Research, Keck School of Medicine of USC

Summary

Multiomics provides a more comprehensive view of complex biological systems by bringing together information across multiple molecular layers. Although multiomic study designs vary widely, all benefit from analytic frameworks that can resolve complex, multilayered data sets with accuracy and consistency. Illumina application-based solutions streamline data analysis and support confident interpretation, helping researchers move efficiently from sequencing data to biological insight. The result is a simpler path from complex multiomic data to new insights and discoveries.

Get started with your multiomic data analysis today

Learn more about Illumina multiomics →



References

1. Ohashi H, Hasegawa M, Wakimoto K, et al. Next-generation technologies for multiomics approaches including interactome sequencing. *Biomed Res Int*. 2015;2015:104209. doi:10.1155/2015/104209
2. Dar MA, Arafah A, Bhat KA, et al. Multiomics technologies: role in disease biomarker discoveries and therapeutics. *Brief Funct Genomics*. 2023;22(2):76-96. doi:10.1093/bfgp/elac017
3. Flynn E, Almonte-Loya A, Fragiadakis GK. Single-cell multiomics. *Annu Rev Biomed Data Sci*. 2023;6:313-337. doi:10.1146/annurevbiodatasci-020422-050645
4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;18(1):83. doi:10.1186/s13059-017-1215-1
5. Digital Science. Dimensions [database]. app.dimensions.ai. 2015–.2025 Accessed December 4, 2025, under license agreement.
6. Illumina. No limits in the cloud: Insights proliferate at Eurofins Viracor. illumina.com/company/news-center/feature-articles/no-limits-in-the-cloud--insights-proliferate-at-eurofins-viracor0.html. Published May 1, 2023. Accessed June 15, 2026.
7. Illumina. Efficient cloud data analysis for COPD multiomics project. illumina.com/science/customer-stories/icommunity-customer-interviews-case-studies/okayama-cloud-data-analysis-copd.html. Published 2023. Accessed June 15, 2026.
8. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat Med*. 2022;28(2):243-250. doi:10.1038/s41591-021-01672-4
9. Miga KH, Wang T. The need for a human pangenome reference sequence. *Annu Rev Genomics Hum Genet*. 2021;22:81-102. doi:10.1146/annurev-genom-120120-081921
10. Wong KHY, Ma W, Wei C-Y, et al. Towards a reference genome that captures global genetic diversity. *Nat Commun*. 2020;11(1):5482. doi:10.1038/s41467-020-19311-w
11. Illumina. The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3. illumina.com/science/genomics-research/articles/second-gen-multigenome-mapping.html. Published August 12, 2024. Accessed June 15, 2026.
12. Zverinova S, Guryev V. Variant calling: Considerations, practices, and developments. *Hum Mutat*. 2022;43(8):976-985. doi:10.1002/humu.24311
13. Illumina, Inc. Data on file, 2026.
14. Koch CM, Chiu SF, Akbarpour M, et al. A beginner's guide to analysis of RNA sequencing data. *Am J Respir Cell Mol Biol*. 2018;59(2):145-157. doi:10.1165/rcmb.2017-0430TR
15. Illumina. Illumina Connected Multiomics. illumina.com/products/by-type/informatics-products/connected-multiomics.html. Published 2026. Accessed June 15, 2026.

illumina®

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

M-AMR-01553 v2.0