Getting started with NGS data analysis

Learn the basics of analyzing NGS data to obtain accurate and high-impact research results

M-AMR-01877

illumina®

Table of contents

Unbiased discovery with NGS answers your boldest research questions

Accurate and comprehensive NGS datasets are helping researchers across the globe to expand the breadth and impact of their research, providing unbiased discovery into biological pathways and systems. Over the last ten years, the costs involved in sequencing, preparation, and analysis have dropped quickly. This has led to a big increase in how many people are using these technologies, allowing more researchers to take advantage of the faster and more efficient benefits of NGS. Consequently, scientists are able to explore a broader landscape of molecular entities (new signaling networks, disease biomarkers, and novel drug targets). NGS analysis lays the groundwork for investigating multiple biological "-omes", such as the genome, epigenome, transcriptome, and proteome, at any biological resolution to explore gene expression and microenvironment differences. You also don't need to be an expert to get started with NGS—local experts and hundreds of core labs across the Americas are available to support you at every step.

Despite the enhanced discovery power of NGS, until recently, the barrier to entry for NGS remained high for researchers in part due to the high costs and expertise required for data analysis and interpretation. Fortunately, commercial tools for data analysis of large NGS data sets have become increasingly accessible, making bioinformatics easier than ever. New, intuitive, user-interface-based tools support every step of analysis—from QC and normalization to advanced statistics, variant interpretation, and visualization—for popular experiments like RNA-Seq and single-cell analysis. You don't need to be an expert to get started with cutting-edge NGS analysis tools. This resource aims to lay the foundation for understanding the key steps for data analysis, plus, how to get started.

Key Terms

Primary analysis, or base calling, is completed automatically on Illumina sequencing systems

Secondary analysis is the post-sequencing process in which sequence transcripts are counted or read through mapping and alignment to a reference genome depending on the application

Tertiary analysis, can involve the analysis and visualization of molecular and clinical research data for cohort analysis, or the annotation, prioritization, and classification of variants depending on the application or automated reporting

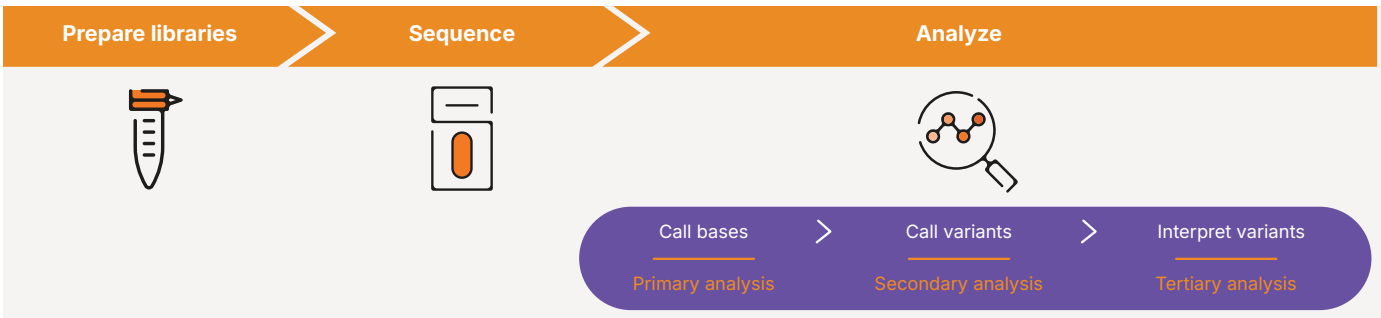


Figure 1: Standard NGS workflow—Library preparation and sequencing are followed by data analysis, which typically includes base calling, variant calling, variant interpretation, and reporting.

Getting started with bioinformatics

While there are many ways to set up your analysis and interpretation workflow, the number of options can sometimes feel overwhelming. Fortunately, NGS analysis typically follows a standardized workflow of three main steps (primary, secondary, tertiary analysis), with room for customization as needed (Figure 1). Understanding your requirements, resources, and goals can help you select analysis tools best suited to your needs. Identifying what matters most to you and your organization in an NGS bioinformatics workflow may include: type of data, expected output, time to result, and budget.

Collaborating with your local service provider or core lab early in the experimental design process is the best way to get expert assistance for your bioinformatics workflow.

Key Terms

Type of data
Volume and size of samples processed, type of files generated post-sequencing, and sensitivity of data (necessary data security and privacy protection)

Expected output
Intended use case and insights derived from the analysis of samples (single sample reporting vs. large-scale cohort analysis)

Time to result
Speed of secondary and tertiary analysis pipelines, report generation, and requirement for automation capabilities

Budget
Costs associated with data analysis computation and storage, institutional guidelines and requirements surrounding data deployment (cloud, on-premises)

Key features to evaluate

Accuracy
High quality insights integrating latest algorithms, databases, and pipelines for evidence backed results

Usability
Intuitiveness and user-friendliness of interface, with easy to use, push-button applications

Support
Continuous support by a team of experts to ensure smooth implementation of analysis workflows

Innovation
Access to the latest technology and robust innovation roadmap to enable the highest quality performance of workflows

Integration
Analysis workflows that work with sequencing, instruments, assays, and existing software to reduce manual touchpoints and data transfer issues

Compatibility
Input for a broad range of application areas, industry standard formats, and protocols to ensure interoperability with other tools and resources

Illumina RNA Workflow Guide
View Here

NGS Analysis for Beginners Page
View Here

Steps in a typical NGS workflow

When analyzing your NGS data, you might hear three terms: primary analysis, secondary analysis and tertiary analysis. What's the difference between these? The NGS data analysis workflow consists of three main steps depicted below.

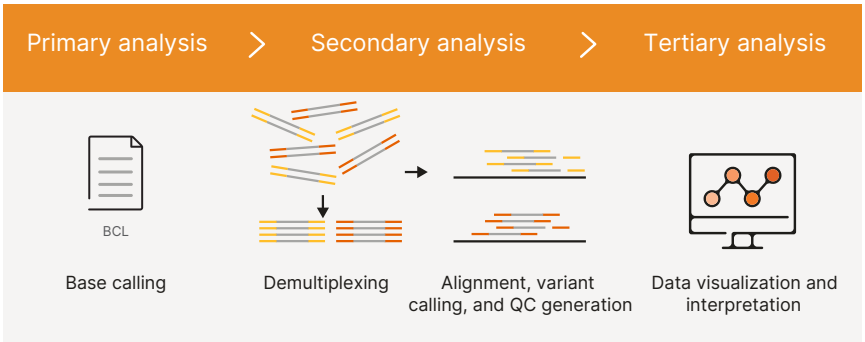


Figure 2: NGS data analysis workflow—NGS data analysis includes three main steps. In primary analysis, raw base call (BCL) files are generated. During secondary analysis, reads are demultiplexed and aligned to a reference genome. The resulting sequence undergoes basic expression profiling. Tertiary analysis involves advanced data visualization and biological interpretation.

Key Terms

Demultiplexing
Demultiplexing in NGS analysis is the process of sorting and assigning sequencing reads to their corresponding samples based on unique barcode sequences

Variant calling
Variant calling is the process of identifying genetic variations, such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), by comparing sequencing reads to a reference genome.

QC generation
QC (quality control) generation in NGS analysis involves assessing sequencing data for quality metrics such as read quality, adapter contamination, GC content, and duplication rates to ensure reliable downstream analysis

Commonly used file formats for Illumina sequencing data

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

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

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

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 that reduces the size and storage cost by 80%, and time to transfer, without compromising data integrity

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

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)

Step 1: Primary analysis

Primary analysis, or Real-Time Analysis (RTA), operates automatically during cycles of sequencing chemistry and imaging, using an on-instrument software. It provides base calls and assigns a quality score to each base.

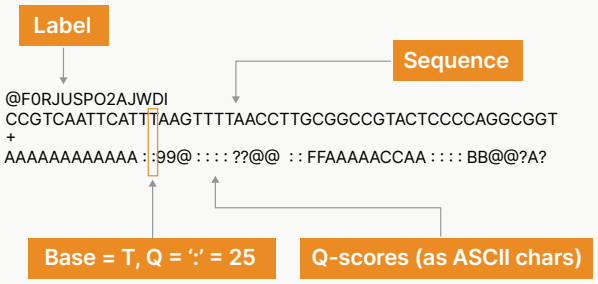


Figure 3: Example FASTQ read with 50 base calls—FASTQ files are text files containing sequence data with a quality (Phred) score for each base, represented as an ASCII character. The quality score is an integer (Q) which is typically in the range 2 - 40, but higher and lower values are sometimes used. 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.

Step 2: Secondary analysis

After sequencing is complete, secondary analysis begins - this can include genetic characterization and expression profiling, *de novo* assembly, protein and epigenetic marker detection, or phasing. BCL files generated during primary analysis are converted into a more versatile, text-based format (FASTQ) after which subsequent steps can occur, such as demultiplexing, mapping, alignment with a reference genome, quality control (QC), gene or transcript quantification, and variant calling.

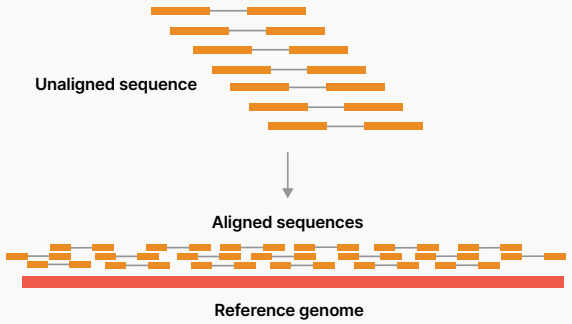


Figure 4: Sequence alignment in secondary analysis—The sequence alignment step maps each base call to the corresponding location on the reference genome or transcriptome. A reference genome is a standardized, generally accepted representation of the genetic sequence of interest. Recent innovations have introduced pangenome or multigenome mapper references, which capture variant diversity of a specific population (based on user input) to significantly improve mapping results and accuracy of the variant calling. The output of this step is a binary alignment map, or BAM, file.

Step 3: Tertiary analysis

The final step of the NGS data analysis workflow is tertiary analysis. It transforms the processed genetic data into meaningful biological insights and can facilitate the transformation of new discoveries into basic biology and disease pathogenesis insights. Key components of tertiary analysis are functional annotation, interpretation, data integration, and reporting.

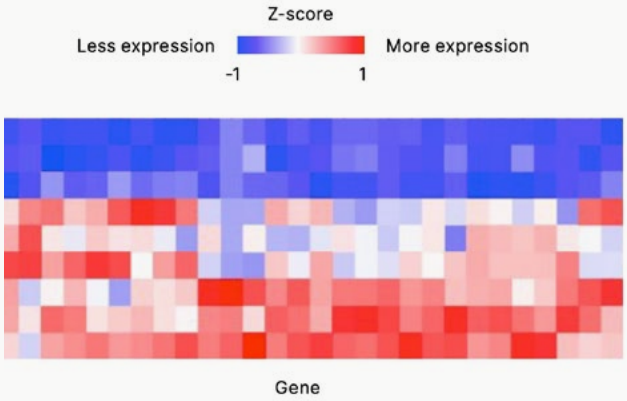


Figure 5: Linking genomics variants with changes in expression—Combined DRAGEN RNA and DNA analysis identifies whether certain mutations correspond to expression patterns in the transcriptome

What Apps can be used to analyze NGS data?

BaseSpace™

BaseSpace is a cloud-based platform for storing, analyzing, and sharing next-generation sequencing (NGS) data. It provides bioinformatics tools, pipelines, and workflows for processing sequencing data, including alignment, variant calling and QC of data

DRAGEN™ Secondary Analysis

DRAGEN provides ultra-fast secondary analysis of data, and can be accessed through a variety of platforms, such as BSSH. For instance, it can analyze single-cell data up to 12X faster than 10X Genomics Cell Ranger and provides unparalleld accuracy

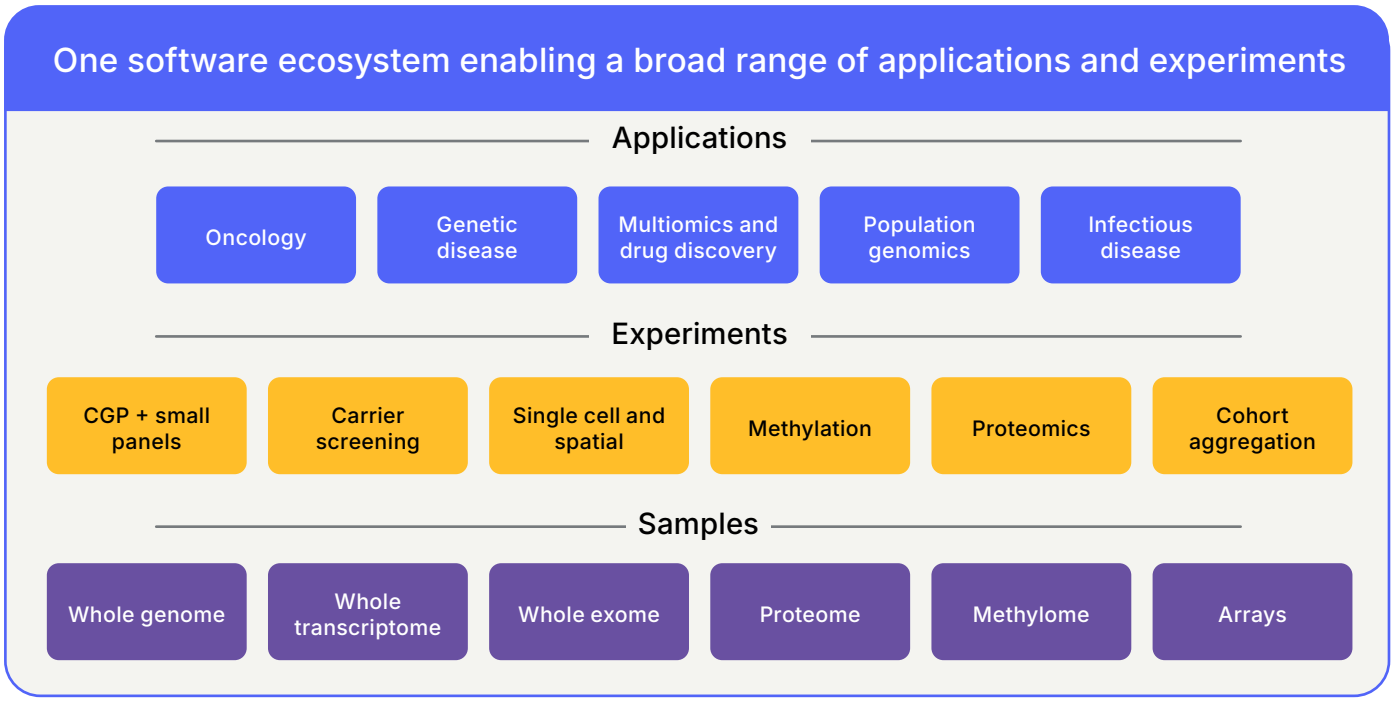
Connected Insights and Emedgene

Connected insights and emedgene support researchers seeking single sample reporting for somatic and germline workflows. Integrated artificial intelligence and customizable reporting provide researchers with easy to read reports they can trust.

Illumina Connected Multiomics

Illumina connected multiomics is Illumina's next-gen tool to analyze single cell, spatial, and multiomic datasets in a user interface-based, intuitive tool. New Illumina multiomic workflows are being designed with end analysis in mind and will be readily analyzed using this platform

The following chart highlights applications and deployment options offered as part of Illumina's data analysis and interpretation solutions.



Your analysis, deployed anywhere

Cloud analysis

Start running pipelines immediately without needing to invest in a data infrastructure. Analysis stored in enterprise-level cloud environments help ensure NGS data is processed and stored securely and privately and makes sharing data easy.

Onboard analysis

The most popular secondary analysis pipelines available on board select Illumina sequencing instruments.

Local analysis

Take maximum control of data storage with on-premises solutions. These solutions require thoughtful planning to balance compute, storage, budget, and security requirements.

When you're studying something that's completely new or uncharacterized, you need a much bigger picture. In our project, there was a lot of uncertainty about what mechanisms are influencing our pathways. And given how little we knew about the mechanisms, we needed to use a broader approach like NGS to more fully understand the mechanism in its entirety.

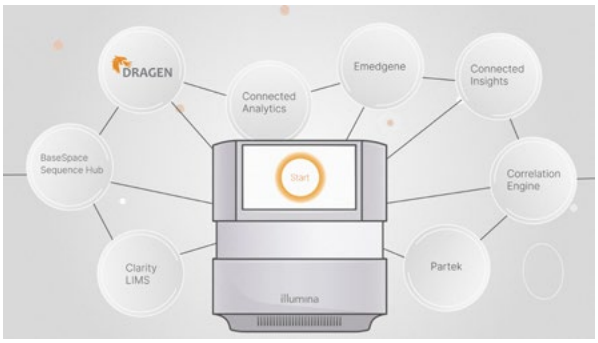
Amanda Touey, PhD candidate in Dr. Paula Cohen's lab at Cornell University

The best time to talk to a bioinformatician is as early as possible. You want to go into your sequencing project knowing what questions you're answering and what analysis you need. And you will only get that by speaking to a bioinformatician or a sequencing core who have some expertise in those things.

Morgan Taschuk, Ontario institute for Cancer Research Director, Genome Sequence Informatics

Learn more about Illumina Connected Software

Illumina software simplifies the process of gaining largescale, unbiased insights from NGS data sets, providing highly accurate and intuitive solutions for every step of the bioinformatics workflow from sample to insights. The Illumina software portfolio features secure, versatile solutions with award-winning accuracy, direct integration with Illumina sequencing systems, and push-button workflows. Illumina software supports a wide range of applications, including various -omes (transcriptomics, proteomics, genomics), to enable biologists in their journey towards next-generation discovery.



1. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions - precision FDA Challenge. Accessed June 7, 2023. https://precision.fda.gov/challenges/10
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

illumina®

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

M-AMR-01877