DATA SHEET

DRAGEN™ secondary analysis

Accurate, comprehensive, and efficient variant calling for next-generation sequencing data

illumina®

Introduction

Unlocking the power of the genome through next-generation sequencing (NGS) is critical to advancing biomedical research and precision medicine. To maximize genetic insights from NGS, researchers require data analysis tools that can accurately and efficiently translate raw sequencing data into meaningful results. Furthermore, to harness the benefits of NGS, organizations require easy-to-use solutions that accommodate a range of users and have lower financial and technical barriers to adoption.

Illumina DRAGEN (Dynamic Read Analysis for GENomics) secondary analysis was developed to address important challenges associated with analyzing NGS data for a wide range of applications, including whole-genome, exome, transcriptome, methylome studies, and more. DRAGEN secondary analysis software is a suite of applications that processes NGS data and enables tertiary analysis to drive insights. The available tools make up a highly accurate, comprehensive, and efficient solution that enables labs of all sizes and disciplines to do more with their genomic data.

Accurate results

DRAGEN secondary analysis generates exceptionally accurate results. In the 2020 Precision FDA Truth Challenge V2 (PrecisionFDA V2), DRAGEN secondary analysis v3.7 won most accurate in all benchmark regions and difficult to map regions with Illumina sequencing data.[1,2] Subsequent releases continue to set new standards in accuracy, with advances in areas including machine learning (ML) and DRAGEN multigenome (graph) technology. The latest version of DRAGEN secondary analysis v4.3 provides unprecedented small variant calling accuracy with a 99.89% F1 score (a combined measure of precision and recall) in all benchmark regions (Figure 1). This is enabled by the third-generation DRAGEN multigenome (graph) reference, built on 128 samples with 256 haplotypes from internally built pangenome reference data, capturing greater genetic diversity. Also contributing to improved accuracy is the new integrated mosaic caller that can be enabled to detect mosaic variants with allele frequencies as low as 3%.

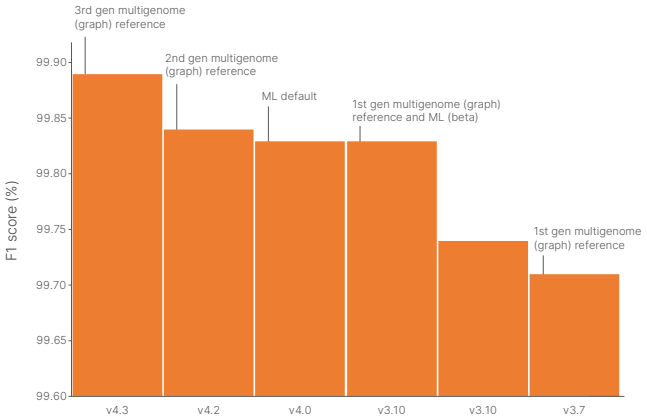

Figure 1: Accuracy of DRAGEN secondary analysis—The Y-axis F1 score (%) is a calculation of true positive and true negative results as a proportion of total results.[3,4]

Comprehensive analysis

With comprehensive coverage of the genome and a broad set of supported applications, DRAGEN secondary analysis meets the diverse needs of labs performing NGS analysis. DRAGEN pipelines support various experiment types, including whole-genome sequencing (WGS), whole-exome sequencing, enrichment panels, single-cell RNA-Seq, single-cell ATAC-seq, bulk RNA-Seq, and methylation analysis (Table 1). It would take over 30 open-source tools to partially replicate the breadth of functionality within DRAGEN software.[3,4]

For germline analysis, DRAGEN secondary analysis includes a suite of variant callers such as ExpansionHunter and targeted callers such as *SMN*, *GBA*, *CYP2B6*, *CYP2D6*, and *HLA*. DRAGEN v4.3 also introduces MRJD, a new specialized caller that enables coverage of difficult genes in segmental duplication regions, such as *PMS2*, *SMN1*, *SMN2*, *STRC*, *NEB*, *TTN*, and *IKBKG*. These tools allow analysis of a broad range of genetic variation, including single-nucleotide variations, insertions and deletions (indels), repeat expansions, and structural variations in extended genomic regions. In addition, DRAGEN multigenome (graph) reference improves mapping quality, leading to greater variant calling accuracy, and resolving areas of the genome that are difficult to assess due to sequence complexities. This increases coverage of potentially medically relevant genes and enables single nucleotide variant, small indel, copy number variation, and structural variant calling in difficult-to-map regions.

Table 1: DRAGEN secondary analysis supports an extensive array of secondary analysis applications[a]

Application	On-premises server	Onboard Illumina sequencing systems		Illumina cloud platforms	
	DRAGEN server	NovaSeq X Series	NextSeq 1000, NextSeq 2000 System	BaseSpace Sequence Hub	Illumina Connected Analytics
BCL convert	✔	✔	✔	✔	Custom only
DRAGEN ORA compression	✔	✔	✔		Custom only
DRAGEN FASTQ + MultiQC	✔	✔	✔	✔	✔
Whole genome	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic
Enrichment (including exome)	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic
DRAGEN Amplicon	✔		DNA only	✔	✔
RNA	✔	✔	✔	✔	✔
Single-cell RNA	✔		✔	✔	✔
NanoString GeoMx NGS			✔	✔	
Methylation	✔	✔		✔	✔
Metagenomics	✔[b]			✔	
RNA pathogen detection				✔	
COVID	COVIDSeq. COVID lineage		COVIDSeq (cloud only)	COVIDSeq. COVID lineage	
TruSight Oncology 500 portfolio	✔			✔[c]	✔
scATAC-seq	✔			✔	✔
Imputation	✔			✔	✔
PGx Star Allele Caller	✔	✔	✔	✔	✔
Illumina Complete Long Reads				✔	✔
DRAGEN secondary analysis for RPIP and UPIP	✔			✔	Beta

a. Core DRAGEN software version varies across platforms, speak to a local representative for more information.
b. Metagenomics applications enabled by Kmer classifier, more tools coming soon.
c. Illumina Connected Analytics subscription required.

Efficient analysis

DRAGEN software is designed to give labs the data analysis speed they need to optimize the efficiency of their NGS data sets processing. DRAGEN secondary analysis is hardware accelerated and uses field-programmable gate array (FPGA) architecture to achieve rapid turnaround times. The efficiency of DRAGEN analysis algorithms resulted in two world speed records for genomic data analysis.[5,6] In practical applications, the on-premises DRAGEN secondary analysis can process NGS data for a whole genome equivalent at 40× coverage in about 35 minutes with all callers* vs. > 8 hours with commonly used open-source methods calling a limited number of variant types.[7]

To make it easier to store, manage, and share large NGS data files, DRAGEN Original Read Archive (ORA) technology provides up to 5× lossless compression of FASTQ files in traditional fastq.gz format. The lossless compression of DRAGEN ORA maintains the details of FASTQ files and is remarkably fast, requiring ~8 minutes for compression of 50–70 GB FASTQ files, supporting a wide range of commonly studied species. DRAGEN secondary analysis features a versatile set of pipelines that can also accept input data files and create output files at different stages of the pipelines (Figure 2).

FPGA and hardware-acceleration

The highly configurable FPGA allows for ultraefficient hardware-accelerated implementations of genomic analysis algorithms, such as base call (BCL) file conversion, mapping, alignment, sorting, duplicate marking, and haplotype variant calling. The flexible nature of FPGAs enables Illumina to develop an extensive suite of DRAGEN application pipelines, with frequent updates and additions to deliver the best possible accuracy, comprehensiveness, and efficiency.

* Based on Illumina internal data based on HG001-HG007 standards on DRAGEN server v4, without new specialized callers like MRJD and VNTR available in DRAGEN v4.3.

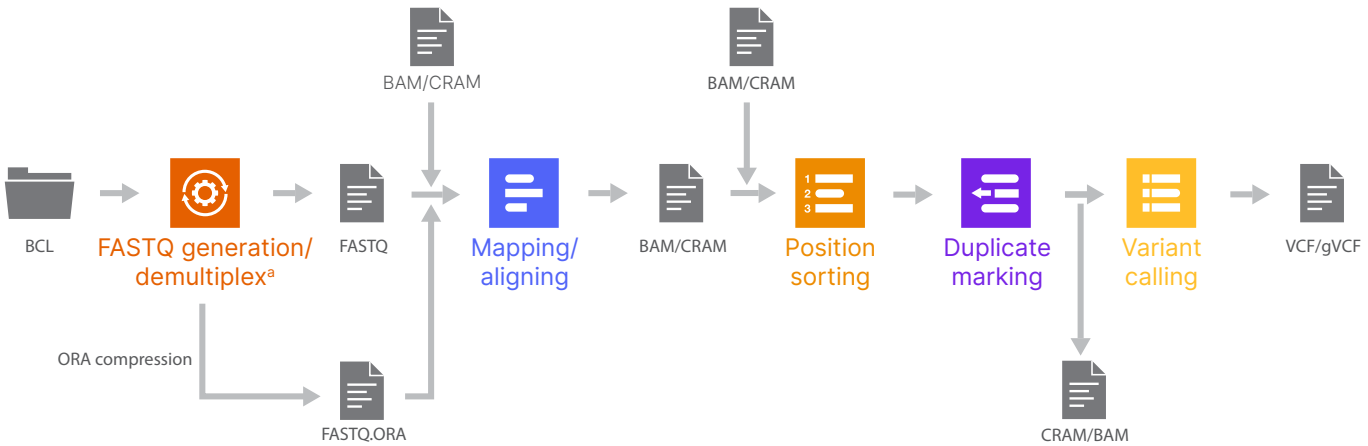

Figure 2: Flexibility of DRAGEN secondary analysis pipelines—Each DRAGEN pipeline contains a specific set of steps to support accurate and efficient analysis. The DRAGEN whole genome germline example pipeline provides the flexibility to accept various input files and produce a range of output types, enabling users to customize their experience and produce their desired file format.

a. BCL convert is also available as a standalone tool.

Custom references

DRAGEN secondary analysis enables users to generate a custom human, nonhuman, or nonstandard reference. Created references can be used as input for all DRAGEN applications that support custom reference files. Most DRAGEN pipelines include built-in support for hg19, hg38 (with or without HLA), GRCh37, CHM13v2, and hs37d5. DRAGEN software enables users to extend graph standard multigenome reference capabilities for both diverse and specific populations.

Scalability

DRAGEN secondary analysis enables labs to scale operations as needed while keeping costs and turnaround times low. DRAGEN software can facilitate the expansion of research capabilities in several ways:

1. **Keeping up with the NovaSeq™ X Series and NextSeq™ 1000 and NextSeq 2000 Systems** — DRAGEN onboard can perform multiple simultaneous applications (four simultaneous applications with a maximum of one BCL convert and three other pipelines of your choice) per flow cell in a single run.

2. **Burst capacity**—During times of increased workloads with high sample volumes, labs can take advantage of additional on-cloud capacity with DRAGEN secondary analysis on Illumina Connected Analytics or DRAGEN apps on BaseSpace Sequence Hub (Figure 3).

3. **Expanding operations**—A single DRAGEN instance can run a broad range of DRAGEN pipelines and supported sample types. The comprehensiveness and efficiency of DRAGEN software enable users to scale up operations without compromising turnaround times or quality of results.

4. **Transition to genomes**—DRAGEN prebuilt pipelines enable easy transition from targeted panels to exomes to genomes.

5. **Large population genomics initiatives**—DRAGEN secondary analysis offers a simplified workflow for large-scale cohort analysis, featuring multiple pipelines that are used in conjunction to call genetic variations with high accuracy. DRAGEN gVCF Genotyper enables aggregation of thousands to millions of genomic variant call format (gVCF) files and incorporates new batches without reprocessing existing batches. ORA compression saves on storage costs.

6. **Deep sequencing applications**—DRAGEN secondary analysis supports analysis of high-depth sequencing with high efficiency for average coverage of over 300× for genomes and 1000× for exomes. The deep sequencing capabilities are valuable for applications such as oncology research and rare genetic disease studies.

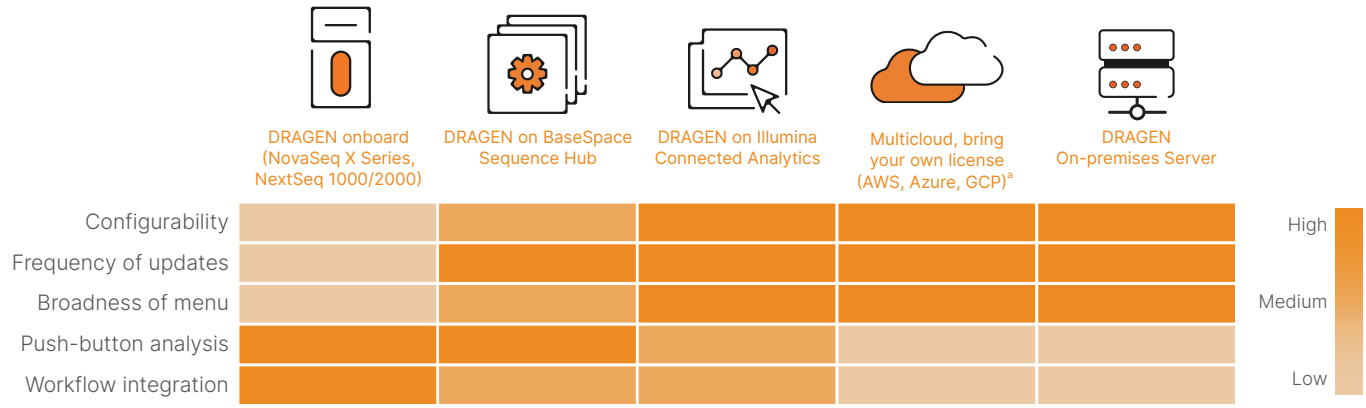

Figure 3: DRAGEN pipeline access options with features designed to fit the NGS analysis needs of every lab.

a. Contact your Illumina representative for information about access on Amazon Web Service (AWS), Azure, or Google Cloud Platform (GCP, early access).

Multiplatform accessibility

The suite of DRAGEN pipelines can be accessed through available on-premises, on-instrument, or cloud solutions, enabling labs to select a solution that best suits their needs (Figure 3).

DRAGEN on-premises server

DRAGEN on-premises server relies on a local storage solution to collect and store NGS data. After raw sequencing data has been transferred from the sequencing instrument to the local storage via a local network connection, it is transferred to the DRAGEN server to perform the selected workflow. Following analysis, the software writes the generated output files back to the local storage location. DRAGEN on-premises server:

- Supports flexible configuration of DRAGEN features through a command-line interface
- Replaces up to 30 traditional compute instances
- Processes NGS data for an entire human genome at 40× coverage in ~35 minutes

DRAGEN onboard the NovaSeq X Series

The NovaSeq X Series includes onboard DRAGEN secondary analysis, offering accurate, automated, and streamlined analysis, designed to support the extraordinary volume of data generated by the NovaSeq X Series. The onboard DRAGEN software suite provides secondary analysis and ORA compression with NGS applications (Table 1) covering BCL convert, germline, somatic, enrichment, RNA, and methylation. DRAGEN onboard:

- Runs multiple secondary analysis pipelines in parallel
- Performs up to four simultaneous applications per flow cell in a single run
- Brings up to 5× lossless data compression and storage cost savings
- Provides savings on analysis, which over five years can exceed the purchase price of the NovaSeq X System

DRAGEN onboard NextSeq 1000 and NextSeq 2000 Systems

NextSeq 1000 and NextSeq 2000 Systems include onboard DRAGEN software for rapid, accurate secondary analysis. The software is accessed through a user-friendly graphical interface that allows expert and nonexpert users to perform needed analyses and produce results quickly. Onboard DRAGEN software offers a select set of pipelines to cover a range of common NGS applications (Table 1) and includes award-winning ML and multigenome (graph) reference analysis for high-quality variant calling. DRAGEN onboard:

- Offers the highest accuracy of any benchtop sequencing system with onboard DRAGEN secondary analysis
- Provides access to select DRAGEN informatics pipelines
- Enables users to generate results in as little as two hours
- Uses intuitive pipeline algorithms to reduce reliance on external informatics experts

BaseSpace Sequence Hub

The cloud-based DRAGEN suite available on BaseSpace Sequence Hub combines accurate, efficient analysis with a secure ecosystem and versatile scalability. DRAGEN software on BaseSpace Sequence Hub enables push-button secondary analysis for labs of all sizes and disciplines. BaseSpace Sequence Hub is a direct extension of your Illumina instruments. Encrypted data flow from the instrument into BaseSpace Sequence Hub, enabling you to manage and analyze your data easily with a curated set of applications. BaseSpace Sequence Hub, powered by Amazon Web Services (AWS):

- Offers a push-button, easy-to-use solution for DRAGEN analysis
- Uses an intuitive graphical user interface for efficient operation by expert and nonexpert users
- Provides access to powerful computing resources without capital expenditure for additional infrastructure

Illumina Connected Analytics

Illumina Connected Analytics is a comprehensive, cloud-based bioinformatics platform that empowers researchers to manage, analyze, and interpret large volumes of multiomic data in a secure, scalable, and flexible environment. Access the DRAGEN secondary analysis suite on Illumina Connected Analytics, available as prepackaged pipelines or individual tools to incorporate into custom pipelines.

Summary

DRAGEN secondary analysis is a powerful suite of software tools that provides accurate, comprehensive, and efficient analysis of NGS data. Multiple DRAGEN software deployment options allow labs to select the solution that best suits the type and scale of their projects. In addition, users can combine various deployment options to best suit their performance and workflow needs. As NGS technology continues to make progress, timely updates to DRAGEN secondary analysis ensure the best possible performance of current pipelines, while new pipelines continue to be added as applications become available.

Learn more

DRAGEN secondary analysis

DRAGEN secondary analysis support page

Contact us

illumina®

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

M-GL-00680 v10.0

References

1. The food and Drug Administration. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. https://precision.fda.gov/challenges/10/results. Accessed March 14, 2022.
2. Catreux S, Jain V, Murray L, et al. DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph. Illumina website. https://www.illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. Published January 12, 2022 Accessed July 11, 2024.
3. Mehio R, Ruehle M, Catreux S, et al. DRAGEN wins at PrecisionFDA Truth Challenge V2 showcase accuracy gains from alt-aware mapping and graph reference genomes. Illumina website. illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html. Published November 9, 2020. Accessed July 11, 2024.
4. Internal data on file. Illumina, Inc., 2022.
5. BioIT World. Children's Hospital Of Philadelphia, Edico Set World Record For Secondary Analysis Speed. bio-itworld.com/news/2017/10/23/children-s-hospital-of-philadelphia-edico-set-world-record-for-secondary-analysis-speed. Published October 23, 2017. Accessed July 11, 2024.
6. San Diego Union Tribune. Rady Children's Institute sets Guinness world record. https://www.sandiegouniontribune.com/2018/02/12/rady-childrens-institute-sets-guinness-world-record/. Published February 12, 2018. Accessed July 11, 2024.
7. Betschart RO, Thiéry A, Aguilera-Garcia D, et al. Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. *Sci Rep*. 2022;12(1):21502. Published 2022 Dec 13. doi:10.1038/s41598-022-26181-3
8. Gross A, Maciuca S, Cox A, et al. Accurate and Efficient Calling of Small and Large Variants from PopGen data sets Using the DRAGEN Bio-IT Platform. Illumina website. www.illumina.com/science/genomics-research/articles/popgen-variant-calling-with-dragen.html. Published May 24, 2021. Accessed March 14, 2022.