DATA SHEET

DRAGEN™ secondary analysis

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

illumina®

Introduction

Unlocking the power of the genome through next-generation sequencing (NGS) is critical to advances in research and 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 range of applications, including genome, exome, transcriptome, and methylome studies. The DRAGEN platform is a secondary analysis software suite 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 v3.7 won most accurate in All Benchmark Regions and Difficult to Map regions for Illumina sequencing data.[1,2] Innovations in Graph Genomes and Illumina Machine Learning (ML) with DRAGEN 4.0 software demonstrates exceptional data accuracy across all sequencing technologies, with an 99.83% F1 score (combined measure of precision and recall) in All Benchmark Regions (Figure 1).[1,2] DRAGEN 4.0 + Graph (ML enabled by default) also has the highest F1 score for most accurate calling compared to all PrecisionFDA V2 submissions in the major histocompatibility complex (MHC) regions.

Comprehensive analysis

DRAGEN secondary analysis meets the needs of labs performing a wide range of NGS applications, providing comprehensive coverage for a broad set of experiment types in a single platform. DRAGEN pipelines support various experiment types, including whole-genome

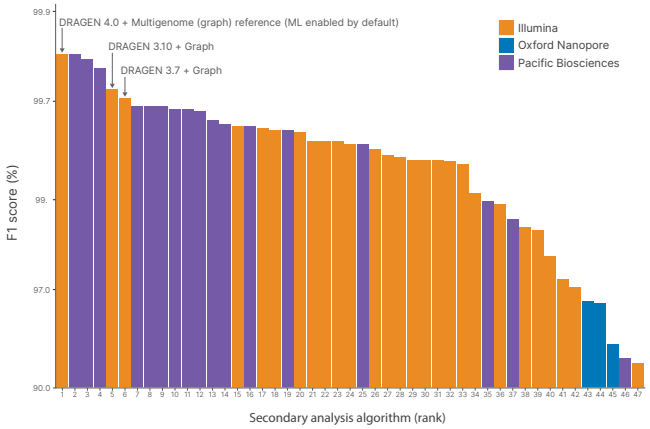

Figure 1: Accuracy of DRAGEN 4.0 + Graph (ML enabled by default) compared to the PrecisionFDA Truth Challenge v2 submissions in the All Benchmark Regions data set—The DRAGEN 4.0 + Graph (ML enabled by default) demonstrates exceptional accuracy, tied with Google DeepVariant on Pacific Biosciences sequencing data. DRAGEN 3.10 + Graph shows improvement over the DRAGEN 3.7 + Graph due to graph and reference/alt-contig handling improvements. The Y-axis, F1 score (%) is a calculation of true positive and true negative results as a proportion of total results.[3,4]

sequencing (WGS), enrichment panels, single-cell RNA-Seq, single-cell ATAC-Seq, single-cell multiomics, 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]

With the included suite of variant callers—repeat expansion, structural variation (SV), copy number variation (CNV), ExpansionHunter, and targeted callers such as *SMN*, *GBA*, *CYP2B6*, *CYP2D6,* and *HLA*—DRAGEN software offers broad genomic coverage. In addition, DRAGEN Multigenome (Graph) reference effectively lengthens Illumina reads and reaches low-complexity regions, resolving areas of the genome that are difficult to assess due to repeat sequences. This improves coverage of potentially medically relevant genes and enables single nucleotide, copy number, and structural variant calling in difficult-to-map regions.

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

Application	DRAGEN On-premise Server	Onboard NovaSeq X Series	Onboard NextSeq 1000, NextSeq 2000 System	BaseSpace Sequence Hub	Illumina Connected Analytics	
					Preconfigured	Customized
BCL conversion	✔	✔	✔	✔	✔	✔
DRAGEN ORA compression	✔	✔	✔			✔
DRAGEN FASTQ + MultiQC	✔	✔	✔	✔	✔	✔
Whole genome	Germline + somatic	Germline only	Germline only	Germline + somatic	Germline + somatic	Germline + somatic
Enrichment (including exome)	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic	Germline + somatic
DNA amplicon	✔		✔	✔	✔	✔
RNA	✔	✔	✔	✔	✔	✔
Single-cell RNA	✔		✔	✔	✔	✔
Differential expression		✔	✔	✔		
NanoString GeoMx NGS				✔		
RNA amplicon	✔			✔	Coming soon	Coming soon
Methylation	✔			✔	✔	✔
Metagenomics				✔		
RNA pathogen detection				✔		
COVID lineage	✔			✔	Coming soon	
TruSight Oncology 500	✔				✔	
ScATAC-Seq	✔			✔	✔	✔
Imputation	✔			✔	✔	✔
PGx Star Allele Caller	✔			✔	✔	✔
Illumina Complete Long Reads				✔		

Efficient analysis

DRAGEN software is specifically designed to give labs the data analysis speed and file options they need to obtain the greatest benefits from NGS data sets. DRAGEN secondary analysis is hardware accelerated and uses field-programmable gate array (FPGA) architechure 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-premise DRAGEN secondary analysis can process NGS data for a whole genome equivalent at 34× coverage in about 30 minutes on-premise vs > 15 hours with a traditional CPU-based system.[7]

To deal with the storage demands of large NGS data files, DRAGEN Original Read Archive (ORA) technology provides lossless 5× compression of FASTQ files. The lossless compression of DRAGEN ORA maintains the details of FASTQ files and is remarkably fast, requiring ~8 minutes for compression of 50 GB to 70 GB FASTQ files. 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.

Custom references

DRAGEN reference builder enables users to generate a human, nonhuman, or nonstandard reference, all referred to as hash tables. Created references can be used as input for all DRAGEN applications that support customer reference files. The DRAGEN Reference Builder application on BaseSpace™ Sequence Hub requires a FASTA file. Most DRAGEN pipelines include built-in support for hg19, hg38 (with or without HLA*), GRCh37, and hs37d5. DRAGEN Graph Toolkit enables users to extend graph reference capabilities to more diverse human graph references as well.

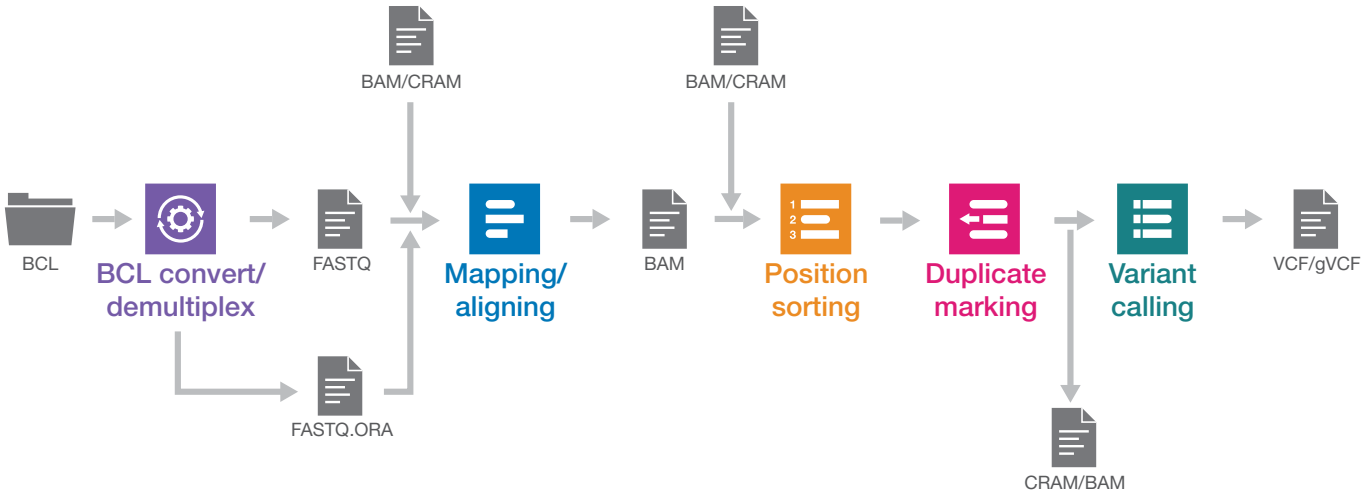

Figure 2: Flexibility of DRAGEN pipelines—Each DRAGEN pipeline contains a specific set of steps to support accurate and efficient analysis. The DRAGEN pipeline provides the flexibility to accept various input files and produce a range of output types, enabling users to customize their experience and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

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**—DRAGEN onboard can perform up to four simultaneous applications 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 capacity through available parallel DRAGEN software access options (Figure 3).

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

4. **Exomes to genomes**—Ramping from whole-exome sequencing (WES) to WGS involves a large increase in generated data. DRAGEN software enables customers to perform analyses from exomes to genomes without large investments in additional hardware infrastructure or cloud-based solutions.

5. **Very large data sets**—DRAGEN secondary analysis offers a simplified workflow for large-scale cohort analysis, featuring multiple pipelines that are used in conjunction to call small and large variants with high accuracy from a cohort sampling. DRAGEN software enables aggregation and genotyping of thousands to millions of genomic variant call format (gVCF) files and aggregates new batches without reprocessing existing batches. The DRAGEN Joint Genotyping Pipeline calls variants jointly across multiple genomes and scales to large cohorts with rapid analysis and uncompromising accuracy.[8] DRAGEN secondary analysis of 1000 Genomes Project data, for example, enabled large scale, accurate variant calling of diverse samples and identification of regions where coverage data are nonuniform or deviate from the assumptions.

Multiplatform accessibility

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

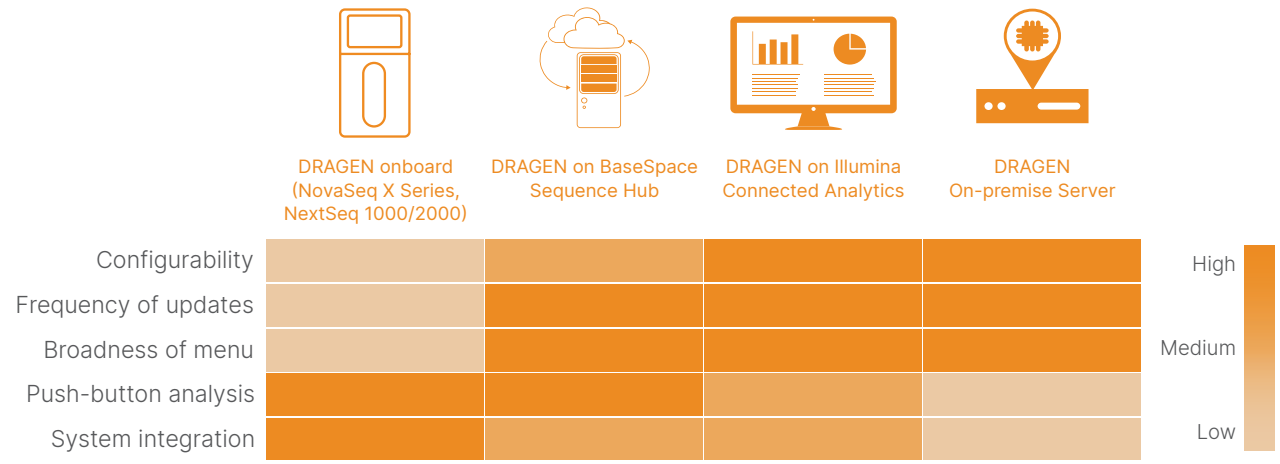

Figure 3: DRAGEN pipeline access options with features designed to fit the NGS analysis needs of every lab.

DRAGEN on-premise server

DRAGEN on-premise 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-premise server:

- Supports varying levels of command-line interface
- Replaces up to 30 traditional compute instances
- Processes NGS data for an entire human genome at 34× coverage in ~30 minutes

DRAGEN onboard the NovaSeq™ X Series

The NovaSeq X Series includes the most powerful DRAGEN software ever, offering accurate, automated, and comprehensive secondary analysis. The onboard DRAGEN software suite provides variant calling and ORA compression with NGS applications (Table 1) covering BCL convert, Germline, RNA, and Enrichment. 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 analysis with supported applications
- 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 onboard DRAGEN software suite offers a select set of pipelines to cover a range of common NGS applications (Table 1) with a user-friendly interface that allows expert and nonexpert users to perform needed analyses and produce results quickly.

DRAGEN onboard:

- 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

DRAGEN secondary analysis on 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. Illumina Connected Analytics:

- Provides access to the complete DRAGEN software, available as prepackaged pipelines or individual tools for custom pipelines
- Supports highly automated workflows and custom solutions for optimized high-throughput studies
- Offers a highly secure environment with guaranteed data residency, single sign-on access, audit logs, and access control supporting Health Insurance Portability and Accountability Act (HIPAA) compliance and

European Union, General Data Protection Regulation (GDPR) principles

Summary

DRAGEN secondary analysis is a powerful suite of software tools that provides accurate, comprehensive, and efficient analysis of NGS data. Multiple DRAGEN software options allow labs to select the solution that best suits the type and scale of their projects. 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

References

1. Food and Drug Administration. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. precision.fda.gov/challenges/10. Accessed March 14, 2022.
2. Illumina. DRAGEN Sets New Standard for Data Accuracy in PrecisionFDA Benchmark Data. Optimizing Variant Calling Performance with Illumina Machine Learning and DRAGEN Graph. illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. Accessed March 14, 2022.
3. Illumina. DRAGEN Wins at PrecisionFDA Truth Challenge V2 Showcase Accuracy Gains from Alt-aware Mapping and Graph Reference Genomes. illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html. Accessed March 14, 2022.
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. Accessed March 14, 2022.
6. San Diego Union Tribune. Rady Children's Institute sets Guinness world record. https://www.sandiegouniontribune.com/95899028-132.html. Published February 12, 2018. Accessed March 14, 2022.
7. 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;7:100. doi: 10.1186/s13073-015-0221-8.
8. Illumina. Accurate and Efficient Calling of Small and Large Variants from PopGen Datasets Using the DRAGEN Bio-IT Platform. www.illumina.com/science/genomics-research/articles/popgen-variant-calling-with-dragen.html. Accessed March 14, 2022.

illumina®

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

M-GL-00680 v5.0