

Germline variant calling accuracy improvements in DRAGEN™ secondary analysis

Optimizing variant calling performance with Illumina machine learning and multigenome mapping



Introduction

Unlocking the power of the genome through next-generation sequencing (NGS) is critical to biomedical research and precision medicine. To maximize insights from NGS, researchers require data analysis tools that can translate raw sequencing data into meaningful results. DRAGEN secondary analysis provides accurate, comprehensive, and efficient secondary analysis of NGS data. Using highly reconfigurable field-programmable gate array (FPGA) technology allows DRAGEN secondary analysis to speed up secondary analysis of NGS data, including mapping, alignment, and variant calling. Additionally, DRAGEN secondary analysis is designed to address common challenges in genomic analysis, such as lengthy compute times, massive volumes of data, and variant calling in challenging genomic regions.

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 against other solutions such as Sentieon, Seven Bridges, and BWA-GATK (Figure 1).^{1,2} In just four years, DRAGEN secondary analysis v4.3 has made significant improvements to this already exceptional performance, providing unprecedented small variant calling accuracy with a 99.89% F1 score, a combined measure of precision and recall, in all benchmark regions with new and impactful features.

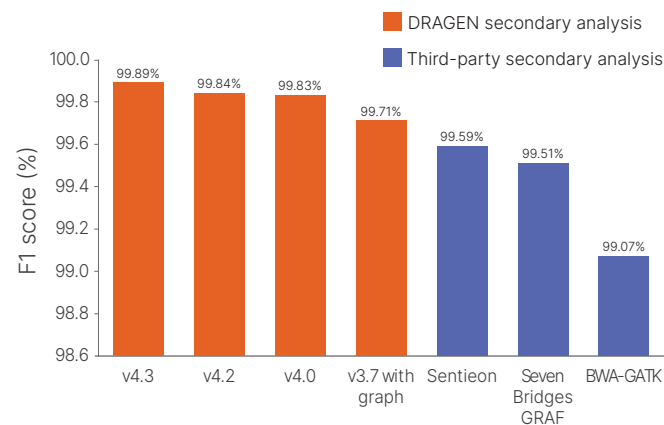


Figure 1: Accuracy of DRAGEN secondary analysis for FDA all benchmark regions analysis—F1 score (%) is a calculation of true positive and true negative results as a proportion of total results.^{5,6} Higher scores indicate improved accuracy based on reference data.

This technical note describes recent improvements that contribute to the high accuracy of DRAGEN secondary analysis, including multigenome mapper with pangenome reference, machine learning (ML) incorporation, mosaic variant calling, specialized callers, and structural variant (SV) and copy number variant (CNV) detection.

Multigenome mapper with pangenome reference

Multigenome mapping, first introduced in DRAGEN secondary analysis v3.7, enables improved variant calling accuracy.³ DRAGEN secondary analysis v4.3 brings significant accuracy gains, with an 83% reduction in errors when compared to v3.6.3 and 40% error reduction when compared to v4.2.7 (Figure 2).

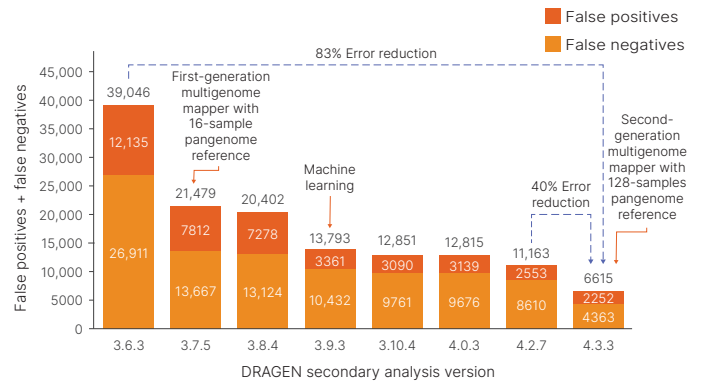


Figure 2: Constant innovation driving DRAGEN secondary analysis—Improvements in false positive and negative rates for SNPs and indels using Genome in a Bottle sample HG002, NIST v4.2.1⁴ demonstrate the significant error reduction that has been achieved in just four years.

To better represent a specific population, DRAGEN secondary analysis v4.3 gives users the option to build a custom pangenome reference, further improving variant calling within their studies. Users can create a custom pangenome reference using their own assemblies or using a selection of assemblies provided by the Human Pangenome Reference Consortium (HPRC). For example, a custom pangenome reference built with 44 HPRC assemblies representing a specific research population yields greater variant calling accuracy in comparison to previous versions of DRAGEN secondary analysis, such as DRAGEN secondary analysis v4.2 (Figure 3). However, the included default pangenome reference (based on 128 samples) in v4.3 should perform best for general use cases.⁴

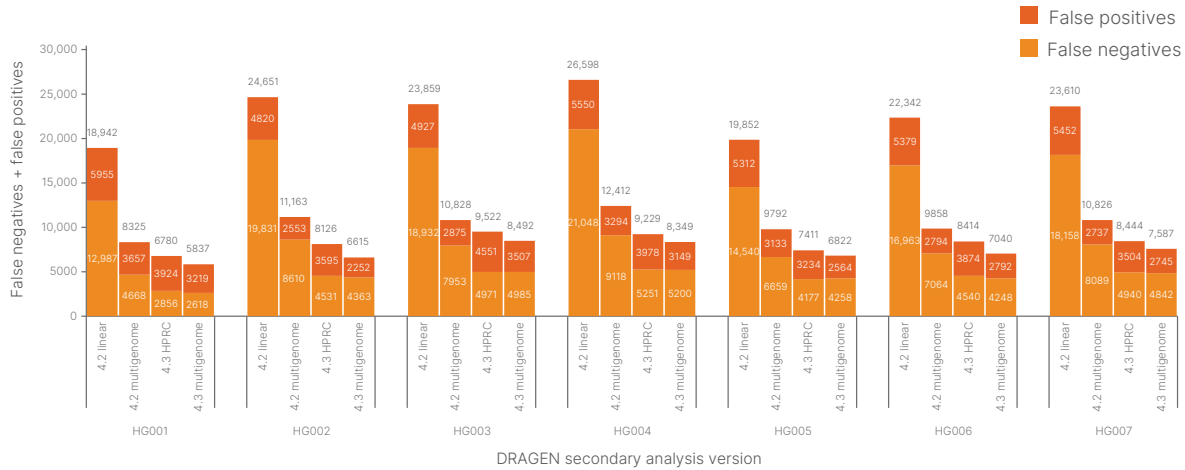


Figure 3: DRAGEN secondary analysis small variant calling accuracy improvements with custom references—DRAGEN secondary analysis v4.3 HPRC-based multigenome reference yields better accuracy results than v4.2 when analyzing Genome in a Bottle samples HG001–HG007.⁴ The default multigenome reference (4.3 multigenome), based on 128 samples, outperforms the 4.3 HPRC-based reference in general use.

Machine learning

The ML module, first added in DRAGEN secondary analysis v3.9 and improved in v3.10, employs a supervised model that uses contextual and read-based features extracted from the DRAGEN secondary analysis variant callers. Small variant accuracy is improved by reducing false calls with the combination of multigenome mapping and ML to deliver the best results (Figure 4). Substantial gains were consistently demonstrated across all subjects, including test data from other populations that were not used during training.

Mosaic variant detection

DRAGEN secondary analysis v4.3, supported by a new ML model, now calls mosaic variants within the germline small variant caller. With the allele frequency threshold decreased to zero, DRAGEN secondary analysis can detect variants with allele frequencies < 20%.

DRAGEN secondary analysis v4.3 detects mosaic variants with greater accuracy and precision than previous versions. To demonstrate this, four DRAGEN secondary analysis pipelines were tested on the National Institute of



Figure 4: ML and multigenome mapping reduce false positives and false negatives—In an analysis of Genome in a Bottle samples HG001–HG007,⁴ ML yields 10% error reduction with multigenome (graph) reference disabled and ~30% error reduction multigenome (graph) reference enabled. When both multigenome reference and ML are enabled, false calls are reduced by 62% for (A) SNVs and (B) indels.

Standards and Technology (NIST) Mosaic truth set data: DRAGEN secondary analysis v4.2, DRAGEN secondary analysis v4.2 in high-sensitivity mode (HSM), DRAGEN secondary analysis v4.3, and DRAGEN secondary analysis v4.3 with Mosaic mode enabled. The NIST Mosaic truth set contains 73 known mosaic variants in 300× data, which were not detected by DRAGEN secondary analysis v4.2 and v4.3, but they were detected by DRAGEN secondary analysis v4.2 in HSM and by DRAGEN secondary analysis v4.3 in Mosaic mode. However, DRAGEN secondary analysis v4.3 in Mosaic mode achieved greater mosaic variant calling accuracy, with 73% fewer false positives than DRAGEN secondary analysis v4.2 in HSM (Figure 5).

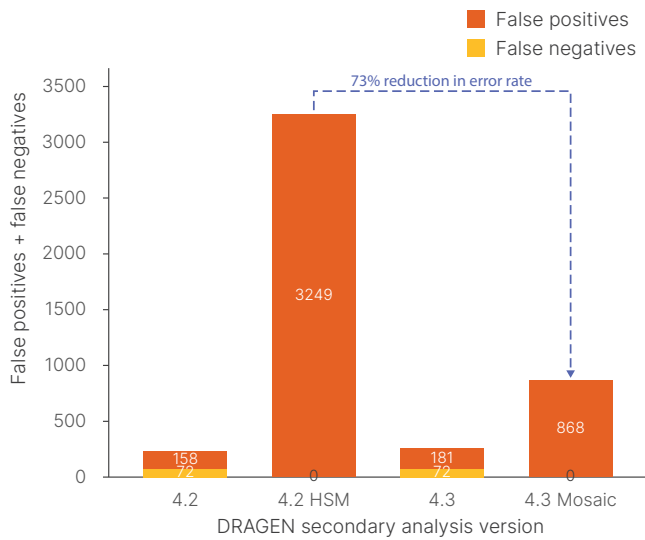


Figure 5: Improved accuracy and precision with mosaic detection mode—There is a 73% error reduction from DRAGEN secondary analysis v4.2 in high sensitivity mode (HSM) compared to DRAGEN v4.3 in Mosaic detection mode. The data also shows the high number of false negatives without HSM mode or mosaic detection enabled.

SV and CNV detection

Structural variants (SV) are genomic alterations that are 50 bp or longer and copy number variants (CNVs) are a specific type of SV where the number of copies of a genomic sequence are reduced (deletions) or increased (insertions). DRAGEN secondary analysis shows greater accuracy for SV calling (Figure 6) and CNV calling (Figure 7) when compared to alternative solutions.⁷ The

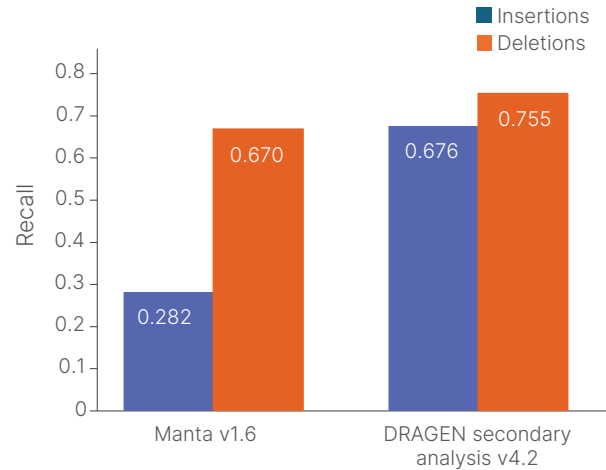


Figure 6: Highly accurate SV calling with DRAGEN secondary analysis—SV indel recall comparison of DRAGEN secondary analysis v4.2 and Manta v1.6 evaluated with Genome in a Bottle (GIAB SV v0.6) benchmark data.⁷

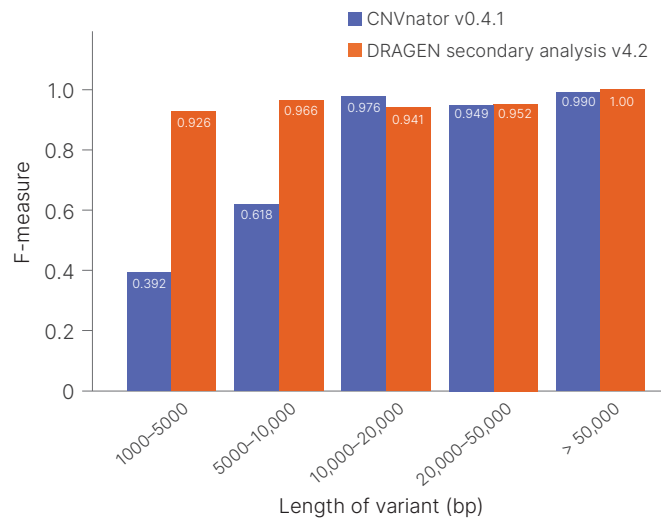


Figure 7: Highly accurate CNV calling with DRAGEN secondary analysis—CNV calling by DRAGEN secondary analysis v4.2 compared to CNVnator v1.6 across different sizes of deletions based on Genome in a Bottle (GIAB SV v0.6) benchmark data.⁷

advanced algorithms and novel approaches tailored for complex genomic regions set DRAGEN secondary analysis apart from other solutions.

The DRAGEN SV caller improves upon the Manta structural variant calling methods and incorporates information from the pangenome reference, leading to more precise filtering

and improved accuracy in SV detection. This includes a new mobile element insertion detector for identifying large insertions, optimized pair parameters for better large deletion calls, and refined contig alignment for improved insertion discovery. Also, DRAGEN software introduces improvements in assembly steps, read likelihood calculations, and improved handling of overlapping mates and clipped bases.

The DRAGEN CNV caller is primarily a read-depth-based caller, with support of various segmentation and scoring models to suit multiple applications. By leveraging additional signal from discordant and split reads, as is done in SV calling, the CNV caller improves sensitivity to capture events as small as 1 kbp.

The DRAGEN CNV caller also has a Segmental Duplication Extension module, a feature that enables CNV detection in segmental duplication regions of the genome. Segmental duplication regions are regions of the genome with > 90% sequence similarity, representing 5% of the genome. These have poor mappability, making variant detection in these regions challenging. The Segmental Duplication Extension rescues approximately one million bases of CNV regions that were previously excluded from analysis. This enables CNV detection across 43 medically relevant genes and improves overall variant calling accuracy.

Specialized and targeted callers

Targeted callers support accurate genotyping of specific genes that are difficult to analyze due to factors such as high sequence similarity to pseudogenes, repetitive regions, and high degrees of polymorphism. DRAGEN secondary analysis addresses these challenges by incorporating various targeted callers (Table 1), enabling precise genotyping of medically relevant genes. For pharmacogenomics (PGx) insights, the PGx Star Allele Caller calls star alleles and metabolizer status for 22 genes (Table 2).

The DRAGEN human leukocyte antigens (HLA) caller enables highly accurate genotyping of HLA class I and II alleles. It aligns reads to a comprehensive database of over 9000 alleles and can aid in applications such as organ transplantation matching, immunogenetics, and disease association studies.

Table 1: Summary of genes addressed by targeted and specialized callers.

Targeted caller	Research application area	Condition association
<i>CYP21A2</i>	Carrier screening	Congenital adrenal hyperplasia (CAH)
<i>HBA</i>	Carrier screening	α -thalassemia
<i>GBA</i>	Carrier screening	Gaucher disease, Parkinson's disease
<i>SMN</i>	Carrier screening	Spinal muscular atrophy
<i>LPA</i>	Cardiovascular disease	Coronary artery disease
<i>RH</i>	Blood typing	–
<i>CYP2B6</i>	PGx	–
<i>CYP2D6</i>	PGx	–
<i>HLA</i>	Transplant matching, immunogenetics	–

Table 2: Genes with PGx relevance addressed by the PGx Star Allele Caller

Gene symbol		
<i>ABCG2</i>	<i>CYP4F2</i>	<i>RYR1</i>
<i>BCHE</i>	<i>DPYD</i>	<i>SLCO1B1</i>
<i>CACNA1S</i>	<i>F5</i>	<i>TPMT</i>
<i>CFTR</i>	<i>G6PD</i>	<i>UGT1A1</i>
<i>CYP2C19</i>	<i>IFNL3</i>	<i>UGTB17</i>
<i>CYP2C9</i>	<i>MT-RNR1</i>	<i>VKORC1</i>
<i>CYP3A4</i>	<i>NAT2</i>	
<i>CYP3A5</i>	<i>NUDT15</i>	

DRAGEN secondary analysis v4.3 introduces a new class of callers that allows for detection of *de novo* variants in regions with segmental duplications. The multiregion joint detection (MRJD) caller implements a haplotype-based *de novo* small variant caller for six medically relevant genes in segmental duplication regions (Table 3).

Table 3: Genes addressed by MJRD caller

Targeted caller	Research application area	Condition association
<i>PMS2</i>	Hereditary cancer screening	Lynch syndrome for colorectal/ endometrial cancer
<i>SMN1</i> (small variants)	Carrier screening	Spinal muscular atrophy
<i>STRC</i>	Carrier screening	Nonsyndromic hearing loss
<i>NEB</i>	Carrier screening	Nemaline myopathy
<i>TTN</i>	Newborn screening, rare diseases	Cardiomyopathy
<i>IKBKG</i>	Newborn screening	Incontinentia pigmenti, hypohidrotic ectodermal dysplasia

Summary


DRAGEN secondary analysis provides highly accurate, comprehensive, and efficient secondary analysis for NGS applications. Continuous improvements provide increased accuracy and expanded coverage of difficult regions of the genome enabling the detection of challenging and medically relevant variants.

Appendix

Multigenome mapping with pangenome reference

By using population haplotypes of phased variants and augmenting the reference index with population-derived alt contigs, DRAGEN secondary analysis can effectively map against a pangenome reference and improve the mapping of Illumina reads in difficult regions. This new feature effectively extends the reach of Illumina reads and enables accurate mapping and variant calling in regions that previously could not be accessed.

A multigenome mapper is an approach to aid mapping with population data where alternate sequence content, observed in the population, is represented as various diverging and converging paths (Figure 8). Sample reads can be aligned to any best-matching path through the multigenome mapper.

 Learn more, [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.](#)

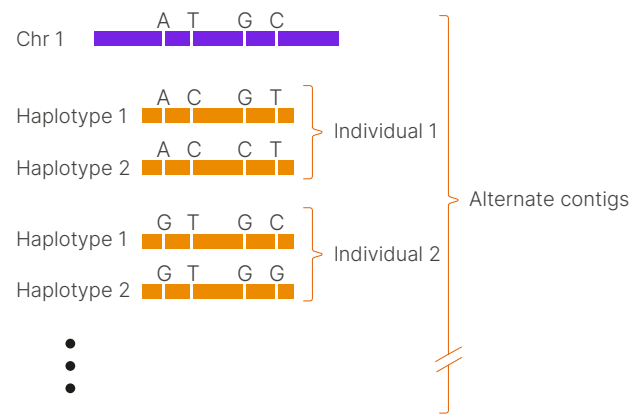



Figure 8: Multigenome mapper with pangenome reference—In a reference, alternate sequence content recorded in a population is represented as various diverging and converging paths.

Alt-masking

Since the DRAGEN secondary analysis v3.9 software update, DRAGEN software includes Alt-masking, a new approach to handle native reference ALT contigs, where strategic positions of the ALT contigs are masked to increase accuracy. This approach is simple to define, maintain, and refine over time.

 Learn more, [DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph](#)

Machine learning

DRAGEN secondary analysis v3.9 software added a powerful and efficient ML recalibration pipeline as an option within the germline small variant workflow. It is enabled by default in DRAGEN secondary analysis v4.0 software. The pipeline runs the ML model after standard variant calling when enabled. This step recalibrates the QUAL and GQ fields that are output to the final VCF. In some cases, ML can change GT. The premachine learning values of these fields are preserved in the DQUAL, DGT, and DGQ fields so that no information is lost. This step adds approximately five minutes to the standard workflow for a 30× WGS germline run so the accuracy improvements have a limited impact on the total run time.


The ML model is generated using supervised offline training. The model processes a set of read-based and contextual features to refine the accuracy of the small variant caller quality scores. The features used to train the model include mappability, AF, VC-Qual, DP, GC content, mismatches and other internal mapping, alignment, and VC metrics.

F1 score computation

$$F1 = 2 \times (\text{Recal} \times \text{Precision}) / (\text{Recall} + \text{Precision})$$

$$F1_{\text{parents}} = \sqrt{F1_{\text{HG003}} \times F1_{\text{HG004}}}$$

DRAGEN command line

 Find starter recipes at [DRAGEN recipe-germline WGS](#)

illumina[®]

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

© 2025 Illumina, Inc. All rights reserved. All trademarks are the property of Illumina, Inc. or their respective owners.
For specific trademark information, see www.illumina.com/company/legal.html.
M-GL-01016 v3.0

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/results. Accessed September 19, 2024.
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. Published January 12, 2022. Accessed September 19, 2024.
3. 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 September 30, 2024.
4. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025. Published 2016 Jun 7. doi:10.1038/sdata.2016.25
5. 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 September 19, 2024.
6. Internal data on file. Illumina, Inc., 2022.
7. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. Published online October 25, 2024. doi:10.1038/s41587-024-02382-1