Germline variant calling accuracy improvements in DRAGEN™ secondary analysis

Optimizing variant calling performance with Illumina machine learning and multigenome mapping

Improved accuracy

Machine learning and multigenome mapper with pangenome reference significantly reduce errors

Mosaic variant detection

Germline small variant caller supported by machine learning calls mosaic variants with accuracy and precision

Expanded coverage

Specialized and targeted callers provide broad coverage of difficult-to-map regions of the genome

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 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] DRAGEN v4.4 has made significant improvements to this already exceptional performance, providing unprecedented small variant calling accuracy with a 99.90% F1 score, a combined measure of precision and recall, in all benchmark regions with new and impactful features. A recent study from the Baylor College of Medicine demonstrated how DRAGEN secondary analysis outperforms current state-of-the-art methods in speed and accuracy across all variant types, including single-nucleotide variations, insertions or deletions, short tandem repeats, structural variations, and copy number variations.[3]

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 improvements to structural variant (SV) and copy number variant (CNV) detection.

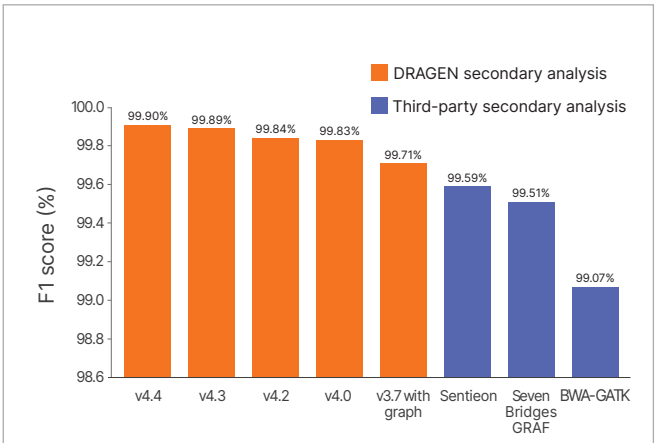

Figure 1: Accuracy of DRAGEN secondary analysis for all benchmark regions

F1 score (%) is a calculation of true positive and true negative results as a proportion of the total results.[5,6] Higher scores indicate improved accuracy based on reference data.

Multigenome mapper with pangenome reference

Multigenome mapping with a pangenome reference, first introduced in DRAGEN v3.7, enables improved variant calling accuracy.[4] DRAGEN v4.3 marks the introduction of the second generation multigenome mapper with the capability to scale up to a pangenome reference representing a larger number of individuals. DRAGEN v4.4 brings significant accuracy gains, with an 87% reduction in errors compared to DRAGEN v3.6.3 and 47% error reduction when compared to DRAGEN v4.2.7 (Figure 2).

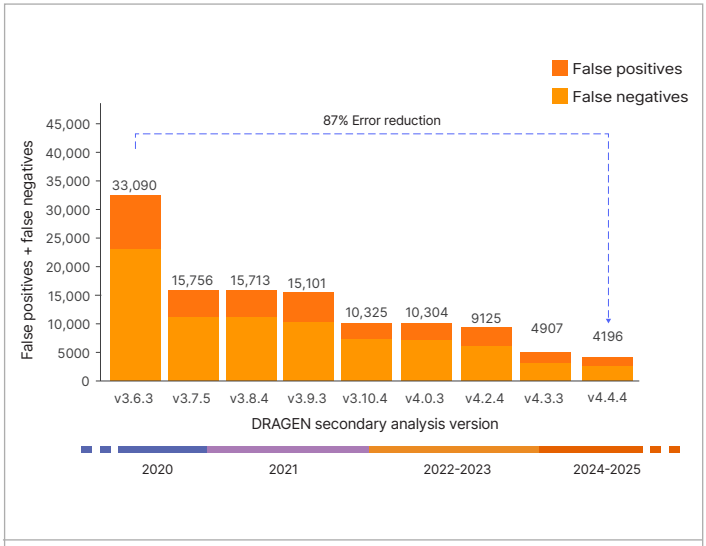

Figure 2: Continuous innovation driving DRAGEN secondary analysis

Improvements in false positive and negative rates for SNPs and indels using Genome in a Bottle sample HG002, NIST v4.2.1[7] demonstrate significant error reduction with superior SNP and indel accuracy in difficult-to-map regions.

Users can create a custom pangenome reference from their own assemblies or using a selection of assemblies provided by the Human Pangenome Reference Consortium (HPRC). Introduced in DRAGEN v4.3, this feature enables users to build references that better represent specific populations, reducing ancestry bias and improving variant calling accuracy. By incorporating population-level haplotypes, the custom pangenome reference enhances mapping performance across diverse cohorts and supports more equitable genomic analysis.

DRAGEN v4.4 introduces the personalized pangenome reference, a distinct feature that tailors the reference to the input sample. Unlike the population-level customization of the custom reference, personalization uses sample-specific data to optimize the reference for each genome. This approach improves mapping accuracy and variant detection, particularly in difficult-to-map regions, with the highest accuracy gains among DRAGEN configurations (Figure 3).

Machine learning

The ML module, first added in DRAGEN 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 2). 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 and later, supported by a new ML model, calls mosaic variants within the germline small variant caller. The mosaic mode can be optionally set to a lower threshold to detect mosaic variants at low allele frequency less than 20% and can be set as low as needed (eg, < 1%) to accommodate high-depth samples. DRAGEN v4.3 and DRAGEN v4.4 detect mosaic variants with greater accuracy and precision than previous versions. To demonstrate this, four DRAGEN analysis pipelines were tested on the National Institute of Standards and Technology (NIST) Mosaic truth set data: DRAGEN v4.2, DRAGEN v4.2 in high-sensitivity mode (HSM), DRAGEN v4.3, and DRAGEN 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 v4.2 and v4.3, but they were detected by DRAGEN v4.2 in HSM and by DRAGEN v4.3 in Mosaic mode. However, DRAGEN v4.3 in Mosaic mode achieved greater mosaic variant calling accuracy, with 73% fewer false positives than DRAGEN v4.2 in HSM (Figure 4).

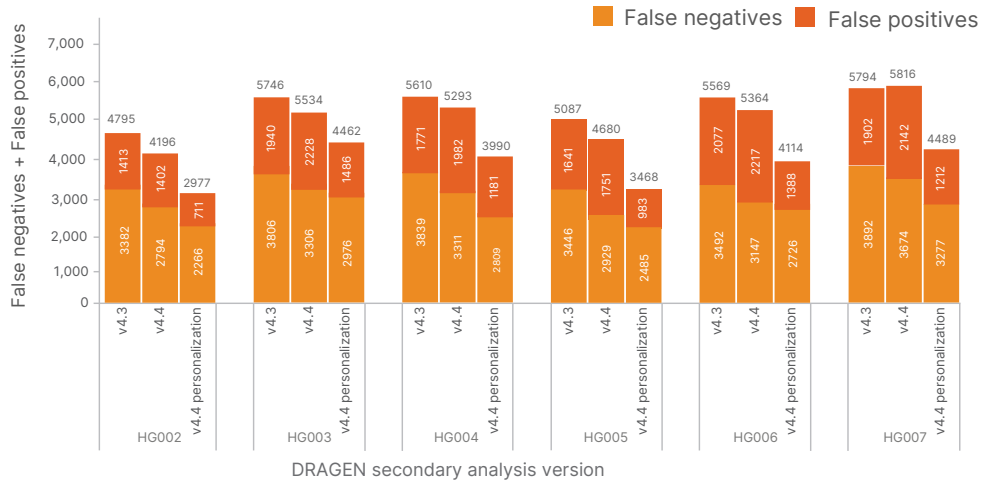

Figure 3: DRAGEN secondary analysis small variant calling accuracy improvements using personalized pangenome references

DRAGEN v4.4 with personalization delivers the highest accuracy among all configurations when evaluated against the NIST v4.2.1 truth sets on difficult-to-map regions. This ability to outperform DRAGEN v4.3 and DRAGEN v4.4 highlights the impact of the personalization feature which was introduced in beta in DRAGEN v4.3 and fully released in DRAGEN v4.4.

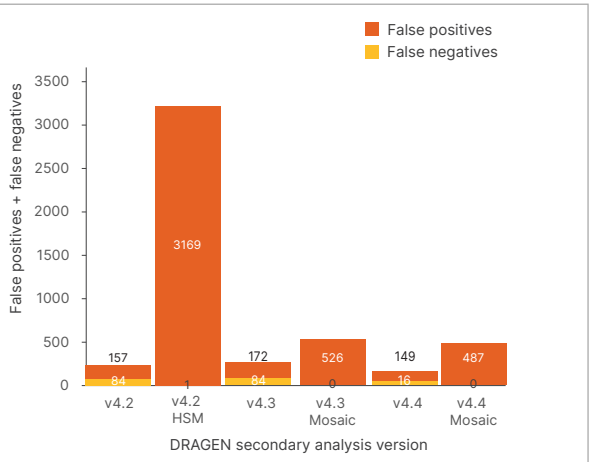

Figure 4: Improved accuracy and precision with mosaic detection mode

There is an ~85% error reduction in DRAGEN v4.4 in Mosaic detection mode compared to DRAGEN v4.2 in high sensitivity mode (HSM). The data also show a higher 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 5) and CNV calling (Figure 6) when compared to alternative solutions.[3] The advanced algorithms and novel approaches tailored for complex genomic regions set DRAGEN secondary analysis apart from other solutions.

DRAGEN v4.4 achieves a major leap in SV calling accuracy by extending the pangenome reference to the SV caller. Structural variant population haplotypes incorporated into the DRAGEN pangenome reference, which includes 128 samples across 26 ancestries. With over a 30% improvement in F1 score, DRAGEN analysis is achieving unprecendeted accuracy for SV calling using short reads. 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 kb. 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. Segmental duplication extension rescues approximately one million bases of CNV regions that were previously excluded from analysis. This module enables CNV detection across 43 medically relevant genes and improves overall variant calling accuracy.

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) for 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.

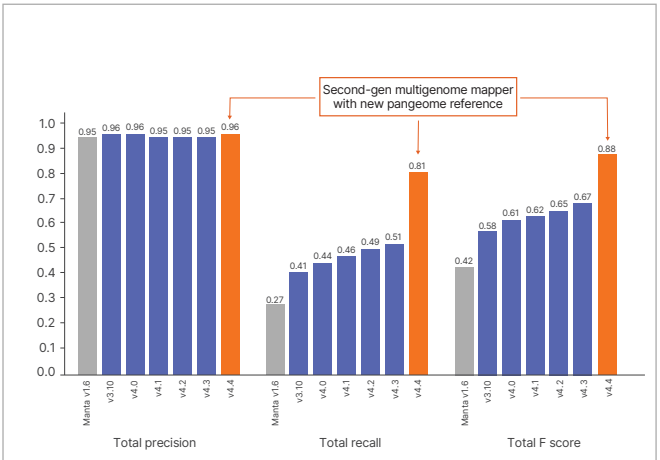

Figure 5: Improved SV calling accuracy with DRAGEN secondary analysis v4.4

SV precision, recall, and F1 score comparison between various DRAGEN versions and Manta v1.6 evaluated with HG002 NIST v1.011 benchmark data.

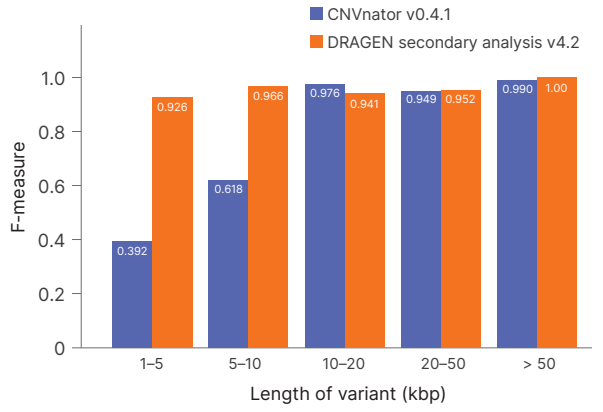

Figure 6: Highly accurate CNV calling with DRAGEN secondary analysis
CNV calling by DRAGEN secondary analysis v4.2 compared to CNVnator v1.6 across different sizes of deletions based on Genome in a Bottle (GIAB SV v0.6) benchmark data.[3]

Table 1: Summary of genes addressed by targeted callers

Targeted caller	Research application	Condition association
CYP21A2	Carrier screening	Congenital adrenal hyperplasia (CAH)
HBA	Carrier screening	α-Thalassemia
GBA	Carrier screening	Gaucher disease, Parkinson's disease
SMN	Carrier screening	Spinal muscular atrophy
LPA	Cardiovascular disease	Coronary artery disease
RH	Blood typing	
CYP2B6	PGx	
CYP2D6	PGx	
HLA	Transplant matching, immunogenetics	Autoimmune disorders, infectious diseases, and certain cancers

DRAGEN secondary analysis v4.3 introduced 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 2: Genes with PGx relevance addressed by the PGx star Allele Caller

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

Table 3: Summary of genes addressed by the MRJD caller

Gene	Research application	Condition association
PMS2	Hereditary cancer screening	Lynch syndrome (colorectal and endometrial cancers)
SMN1, SMN2	Carrier screening	Spinal muscular atrophy
STRC	Carrier screening	Nonsyndromic hearing loss
NEB	Carrier screening	Nemaline myopathy
TTN	Newborn screening and rare disease, ACMG secondary findings list	Cardiomyopathy
IKBKG	Newborn screening	Incontinentia pigmenti, hypohidrotic ectodermal dysplasia

ACMG, American College of Medical Genetics and Genomics.

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.

Learn more →

DRAGEN secondary analysis

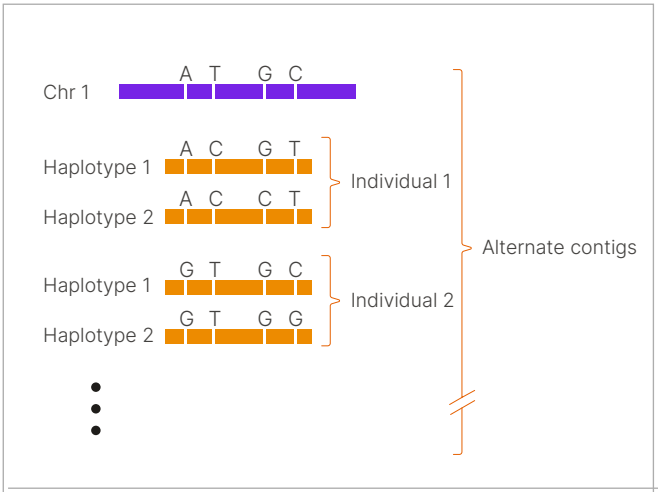

Figure 7: Multigenome mapper with pangenome reference

In a pangenome reference, alternate sequence content recorded in a population is represented as various diverging and converging paths.

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 7). Sample reads can be aligned to any best-matching path through the multigenome mapper.

Read the article: 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.

Alt-masking

Since the DRAGEN v3.9 software update, DRAGEN software includes Alt-masking, an approach for handling 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.

Read the article: 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 and later versions. 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

$$F1 = 2 \times (Recall \times Precision)/(Recall + Precision)$$

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

DRAGEN command line

Find starter recipes at DRAGEN recipe-germline WGS

AF, allele frequency; DP, depth of coverage; GQ, Phred-scaled probability that the call is incorrect; GT, genotyping; QUAL, Phred-scaled probability that the site has no variant; VC-Qual; variant confidence quality.

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 April 3, 2025.
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 April 3, 2025.
3. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. 2024. Published online ahead of print. doi:10.1038/s41587-024-02382-1
4. 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-genmultigenome-mapping.html. Published August 12, 2024. Accessed September 30, 2024.
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 April 3, 2025.
6. Internal data on file. Illumina, Inc., 2022.
7. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025. doi:10.1038/sdata.2016.25

illumina®

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

M-GL-01016 v4.0