TECHNICAL NOTE

Accuracy improvements in germline small variant calling with DRAGEN™ secondary analysis

Optimizing variant calling performance with Illumina machine learning and Multigenome (graph) reference

illumina®

Introduction

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

This application note describes recent advancements in the accuracy of germline small variant calling by the DRAGEN platform, as measured against a truth set. DRAGEN v4.0 software is compared against DRAGEN v3.7 and BWA GATK on Illumina sequencing data and Google DeepVariant on Pacific Bioscience sequencing data, among other various entries in the PrecisionFDA Truth Challenge V2 (Figure 1).

DRAGEN accuracy improvements

Three key DRAGEN innovations drive accuracy improvements over a large portion of the human genome across a wide population of samples:

- Multigenome (graph) reference to improve mapping accuracy in difficult regions
- Alt-masking to reduce mapping ambiguity
- Machine learning to refine small variant calling

Methods

The PrecisionFDA Truth Challenge V2 was sponsored by PrecisionFDA, the Genome in a Bottle (GIAB) consortium, and the National Institute of Standards and Technology (NIST). This challenge was launched to assess small variant calling pipeline performance on a common frame of reference, with a focus on benchmarking in "difficult-to-map" regions, segmental duplications, and the Major Histocompatibility Complex (MHC).

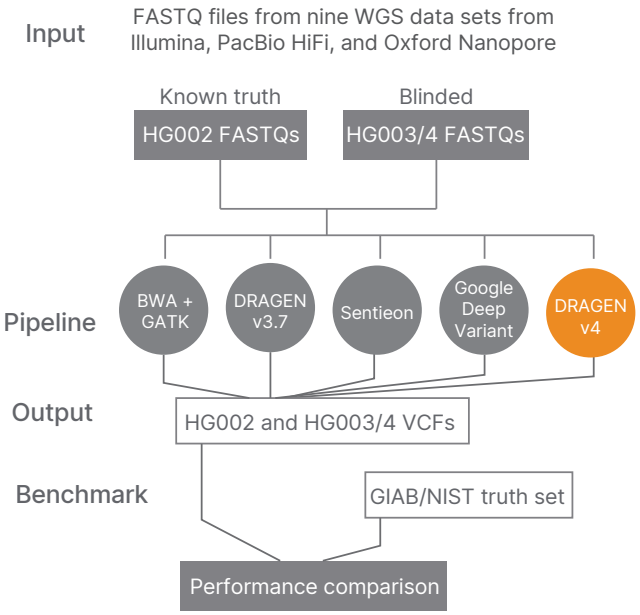

Figure 1: PrecisionFDA Truth Challenge overview—FASTQ files from nine data sets (three samples HG002, HG003, HG004) sequenced across three different technologies (Illumina, PacBio, Oxford Nanopore) were run through multiple analysis pipelines to generate query VCF files that were compared against GIAB/NIST truth sets.

The GIAB consortium used linked and long reads to develop an expanded set of high-confidence truth calls that now covers 7% more of the genome compared to earlier truth sets, including many medically relevant genes. The extended truth set covers over 270 million bases in low-mappability regions and segmental duplications.[1]

These new truth sets were the basis for a new bioinformatics challenge to determine who could demonstrate the best methods for mapping and calling variants in these difficult regions from limited sample data across three different sequencing technologies: Illumina, Pacific Bioscience, and Oxford Nanopore (Figure 1). The PrecisionFDA Truth Challenge V2 samples, truth sets, and results were used to benchmark recent advancements in DRAGEN v4.0.

Results

DRAGEN software leads accuracy in all benchmark regions

DRAGEN secondary analysis generates exceptionally accurate results. In the 2020 Precision FDA Truth Challenge V2, DRAGEN v3.7 won most accurate in All Benchmark Regions and Difficult to Map regions for Illumina sequencing data.

Since the PrecisionFDA Truth Challenge, continuous innovations in multigenome (graph) reference and Illumina machine learning have enabled DRAGEN v4.0 to set a new standard for data accuracy for single nucleotide polymorphism (SNP) and insertion/deletion (indel) calling across all sequencing technologies, achieving a 99.83% F1 score in All Benchmark Regions (Figure 2, Table 1).

DRAGEN accuracy evaluated by FP+FN

Although DRAGEN v3.7 was already competitive with industry-leading informatics solutions, DRAGEN v4.0 has several new modifications (Appendix) that result in significant accuracy improvements. The results of this benchmarking comparison also demonstrate that DRAGEN v4.0 improvements lead to the highest accuracy across popular analysis pipelines and sequencing technologies for all accuracy metrics analyzed in the study (Figure 2).

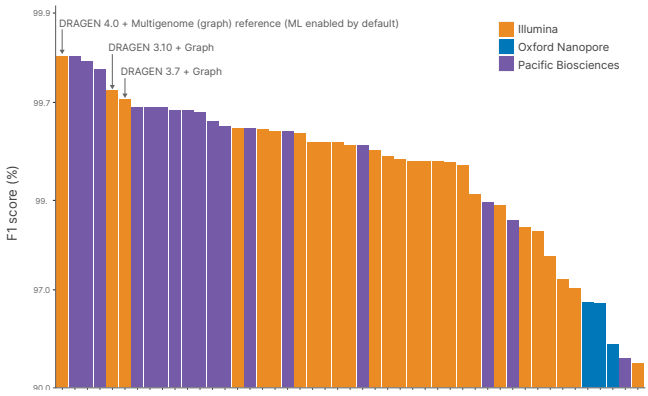

Figure 2: DRAGEN software sets new standards for accuracy—DRAGEN v4.0 +Mulitgenome (graph) reference (Machine learning–enabled by default) demonstrates exceptional accuracy, as measured by F1 score (%), a calculation of true positive and true negative results as a proportion of total results.

Table 1: PrecisionFDA Truth Challenge V2 benchmarking results

Rank	Sequencing technology	Bioinformatics pipeline	F1
1	Illumina	DRAGEN v4.0 with graph (ML default)[a]	0.9983
1	Pacific Bioscience	Google DeepVariant	0.9983
6	Illumina	DRAGEN v3.10 with graph[a]	0.9974
7	Illumina	DRAGEN v4.0 with graph[a]	0.9974
8	Illumina	DRAGEN v3.7 with graph	0.9971
13	Pacific Bioscience	Sentieon	0.9967
17	Illumina	Sentieon	0.9959
28	Illumina	Seven Bridges GRAF	0.9951
30	Pacific Bioscience	Roche	0.9949
39	Illumina	BWA-GATK (Genetalks)	0.9907

a. Not submitted as part of the PrecisionFDA Truth Challenge V2.

To evaluate the impact of the multigenome (graph) reference and machine learning on overall accuracy, we tabulated the false positives (FP) and false negatives (FN) with and without the multigenome reference enabled and with and without machine learning enabled across a range of benchmarking samples (Figure 3). Machine learning yields 10% and ~30% error reduction with multigenome (graph) reference disabled and enabled, respectively. Enabling both multigenome (graph) reference and machine learning has a synergistic effect and reduces false calls by 62%.

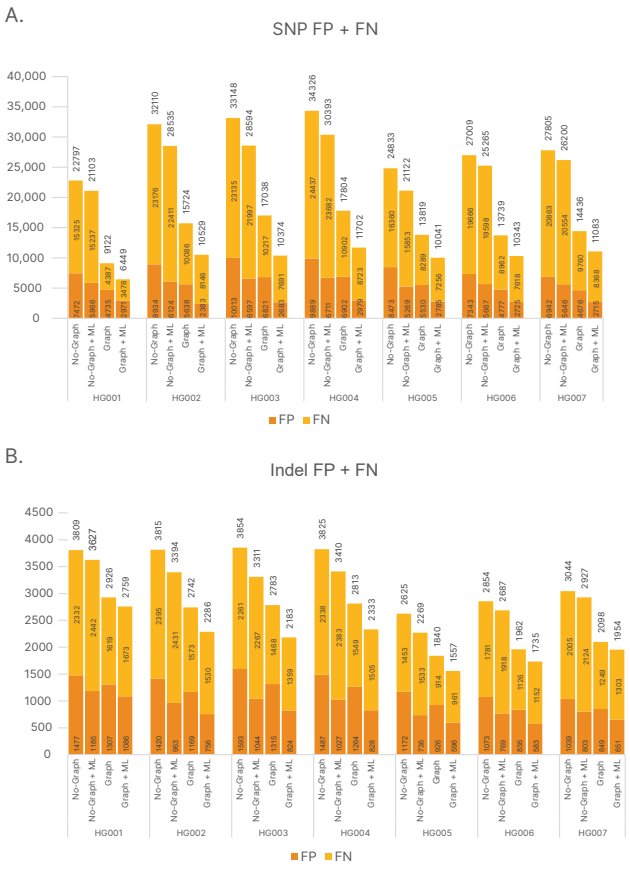

Figure 3: Machine learning and multigenome (graph) reference reduce false positives and false negatives—Machine learning yields 10% and ~30% error reduction with multigenome (graph) reference diasbled and enabled respectively. With both multigenome reference and machine learning enabled, false calls are reduced by 62% for (A) SNVs and (B) indels.

High accuracy in challenging regions

In addition to its strong performance in All Benchmark Regions, DRAGEN v4.0 software demonstrates exceptional accuracy in calling SNPs and indels in the particularly challenging MHC region (Figure 4). The combination of accuracy, comprehensiveness, and efficiency enables DRAGEN customers to unlock the full potential of NGS to maximize genomic insights.

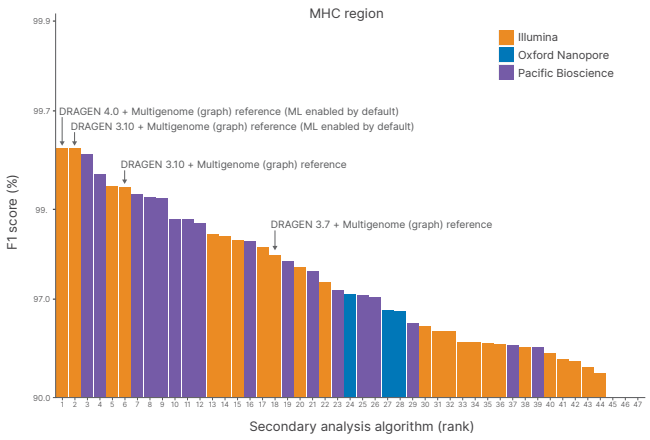

Figure 4: Accuracy of DRAGEN v4.0 in MHC region — DRAGEN v4.0 demonstrates superior accuracy (as measured by F1 score) in the MHC region compared to the PrecisionFDA Truth Challenge V2 entries.

Conclusions

Recent advancements in DRAGEN mapping and germline small variant calling have propelled Illumina sequencing technology with DRAGEN analysis to lead in accuracy across all sequencing and analysis methods compared. DRAGEN's accuracy, combined with its efficiency and comprehensiveness, enables customers to unlock the full potential of genomics.

Summary

DRAGEN secdondary analysis provides highly accurate, comprehensive, and efficient secondary analysis at scale. Continuous accuracy improvements and expanded coverage into the difficult regions of the genome are critical assets for a comprehensive genomic solution. It enables the detection of challenging and medically relevant variants. Improvements in DRAGEN v4.0 yield the most accurate small variant calling across sequencing technologies and analysis methods submitted to the PrecisionFDA Truth Challenge V2.

Appendix

Multigenome (graph) reference

By leveraging population haplotypes of phased variants and augmenting the reference index with population-derived alt contigs, the DRAGEN platform can effectively map against a multigenome (graph) 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 could not be accessed before.

The PrecisionFDA Truth Challenge V2 focused on difficult-to-map regions, the primary regions where the GIAB Consortium expanded their benchmarks. High-quality variant calling with short-read data can be challenging and error-prone in these regions, due to difficulty in mapping short reads to these regions accurately.

Variant callers analyze the pileup of reads mapped to a given locus to determine the most probable original sequence. Mapping difficulty can occur when a region is:

- Highly polymorphic and sample reads differ significantly from the reference genome
- Highly repetitive or contains segmental duplicates and sample reads match reasonably well but with low specificity

A multigenome (graph) reference 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 5). Sample reads can be aligned to any best-matching path through the multigenome reference.

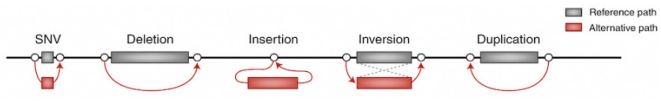

Figure 5: Multigenome (graph) reference representation—In a reference, alternate sequence content observed in a population is represented as various diverging and converging paths.

The DRAGEN platform uses a linear reference as a baseline for read mapping with two capabilities to support augmenting the linear reference into an effective graph reference for highly accurate read mapping in difficult regions:

- Multibase IUPAC-IUB code support for known SNVs (within a population) in the linear reference
- Advanced "alt awareness" capability for alt contigs (alternate paths in a graph), each with predefined liftover alignments into the linear reference

Learn more, illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html

Alt-masking

Since the DRAGEN 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, illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html

Machine learning

DRAGEN v3.9 software added a powerful and efficient machine learning recalibration pipeline as an option within the germline small variant workflow. It is enabled by default in DRAGEN v4.0 software. The pipeline runs the machine learning 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, machine learning can change GT. The pre-machine learning values of these fields are preserved in the DQUAL, DGT, and DGQ fields so that no information is lost.

This step only adds about five minutes for a 30× WGS germline run to the standard workflow, so the accuracy improvements have a limited impact on the total run time.

The machine learning 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.

Detailed methods

Input data sets

Nine preexisting data sets from the PrecisionFDA Truth Challenge V2 across three sequencing technologies (Illumina, Pacific Bioscience, and Oxford Nanopore) and three individuals were leveraged for this analysis. All data sets are publicly available on the PrecisionFDA website.

F1 score computation and benchmark description

We obtained the benchmark values using Wittyer on sample HG003 and HG004 on V4.2 benchmark truth set. The final submission results were evaluated using the geometric mean of the HG003 and HG004 combined SNVs and INDES F1 scores. Specifically,

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

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

DRAGEN command line

```
/opt/edico/bin/dragen
        --output-directory=/output_folder/
        --events-log-file=/output_folder/events_log.csv
        --output-file-prefix=prefix_string
pipeline generated files
        --fastq-file1=/data_folder/fastq_file_1.fastq.gz
        --fastq-file2=/data_folder/fastq_file_1.fastq.gz
        --RGID=DRAGEN_RGID
        --RGSM=read_group_sample_name
        --ref-dir=/reference_genomes/ref_genome
//Initiate DRAGEN
//Specifies the output directory
//Specify log file location
//Outputs file name prefix for all
//Input FASTQ file 1
//Input FASTQ file 2
//Specifies read group ID
//Specifies read group sample name
//Specifies the directory containing the reference hash table
--enablehttp-server=true
--enable-metrics-json=true
--generate-sa-tags=true
--vc-enable-profile-stats=true
--enable-vcf-compression=true
--enable-save-bed-file=true
--enable-variant-caller=true
--enable-map-align=true
--enable-map-align-output=true
--enable-sort=true
--enable-duplicate-marking=true
```

Learn more

DRAGEN secondary analysis, illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html

illumina®

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

M-GL-01016 v2.0