

Population genetics data processing with the DRAGEN™ Bio-IT Platform

Recommendations for data
analysis and variant calling in
large cohorts



Introduction

As the cost of whole-exome sequencing (WES) and whole-genome sequencing (WGS) and downstream data processing continues to decrease, population sequencing studies are becoming feasible at unprecedented scales. Cohort-level catalogs of variation are key resources for ancestry studies, rare variant insights, discovery of genotype/phenotype associations, and annotation of clinical genomic features. Therefore, it is important that cohort call sets are highly accurate, yet informatic and analytical challenges remain when combining data from a large number of samples.

Population genetics data analysis

A typical workflow for population genetics (PopGen) data processing starts with analyzing the samples independently during the read mapping and variant calling stage, with variants exported to a gVCF file. Then gVCF files are aggregated across all samples in a cohort to obtain a conceptual matrix, populated with genotypes and associated confidence metrics (Figure 1). The matrix can be saved in multiple formats, including: a multisample VCF (DRAGEN gVCF Genotyper), multisample gVCF (DRAGEN/Genome analysis toolkit (GATK) Combine gVCF), or a database (GATK GenomicsDB, GLnexus RocksDB). In all cases, the aim is to provide a variant-centric view with genotype calls across the entire cohort. This provides the opportunity to use cohort information to improve genotype calls in individual samples, a statistical model known as joint genotyping. However, care must be taken because increasing sample sizes can also accumulate errors.

There is limited data on the impact of joint genotyping on accuracy, in part because it has been difficult to separate the joint genotyping tool from the gVCF aggregation tool. Aggregating a large number of samples presents particular challenges when unifying different variant representations in a consistent way across the cohort. An increase in cohort size implies an increase of multiallelic variants and the number of alternative alleles, so trade-offs must be made between preserving the full data from the gVCFs and scalability. Additionally, established GATK workflows for data processing are complicated adding to the challenge.

The DRAGEN Platform offers a simplified workflow for cohort analysis (Figure 1) where the output format before and after joint genotyping is multisample VCF file. This enables a direct measurement of the impact of the joint genotyping model.

In this technical note, the performance of joint genotyping with the DRAGEN Platform is evaluated in three use cases that are common for large-scale PopGen projects:

- High-coverage WGS samples at 35×
- Low-coverage WGS samples at 15×
- High-coverage WES samples at 50×

Benchmarking comparisons using the DRAGEN Platform against a call set generated with GATK on a recent resequencing of the 1000 Genomes Project phase 3 samples¹ are presented. Contribution of each workflow stage to call set accuracy is analyzed and a detailed investigation into why some methods that are part of the GATK best practice workflow are not expected to be beneficial for DRAGEN-generated data is provided. Finally, recommendations for processing cohorts with the DRAGEN Platform to obtain analysis-ready variants are presented.

Methods

Input data sets

WGS cohort analysis was based on the 1000 Genomes Project cohort.² The data set contains 2504 WGS samples sequenced using the NovaSeq™ 6000 System at > 30× coverage. Results from processing the same samples with the GATK workflow are publicly available, so results can be reproduced.^{3,4} WES cohort analysis was based on a panel of 10 samples comprising eight unrelated samples from the CEPH collection (CEU) and two samples from trios in the Genome In a Bottle (GIAB) consortium.⁵ All samples were sequenced using the NovaSeq™ 6000 System. The human reference genome hg38 with alternate contigs was used for all analyses.

Cohort analysis

For WGS analysis, gVCFs from the cohort were aggregated and joint genotyped using the DRAGEN Platform v3.5.7b or joint genotyped and processed for variant quality score recalibration (VQSR) following the GATK v3.5 workflow (Figure 1). Both workflows produce msVCF output per chromosome.

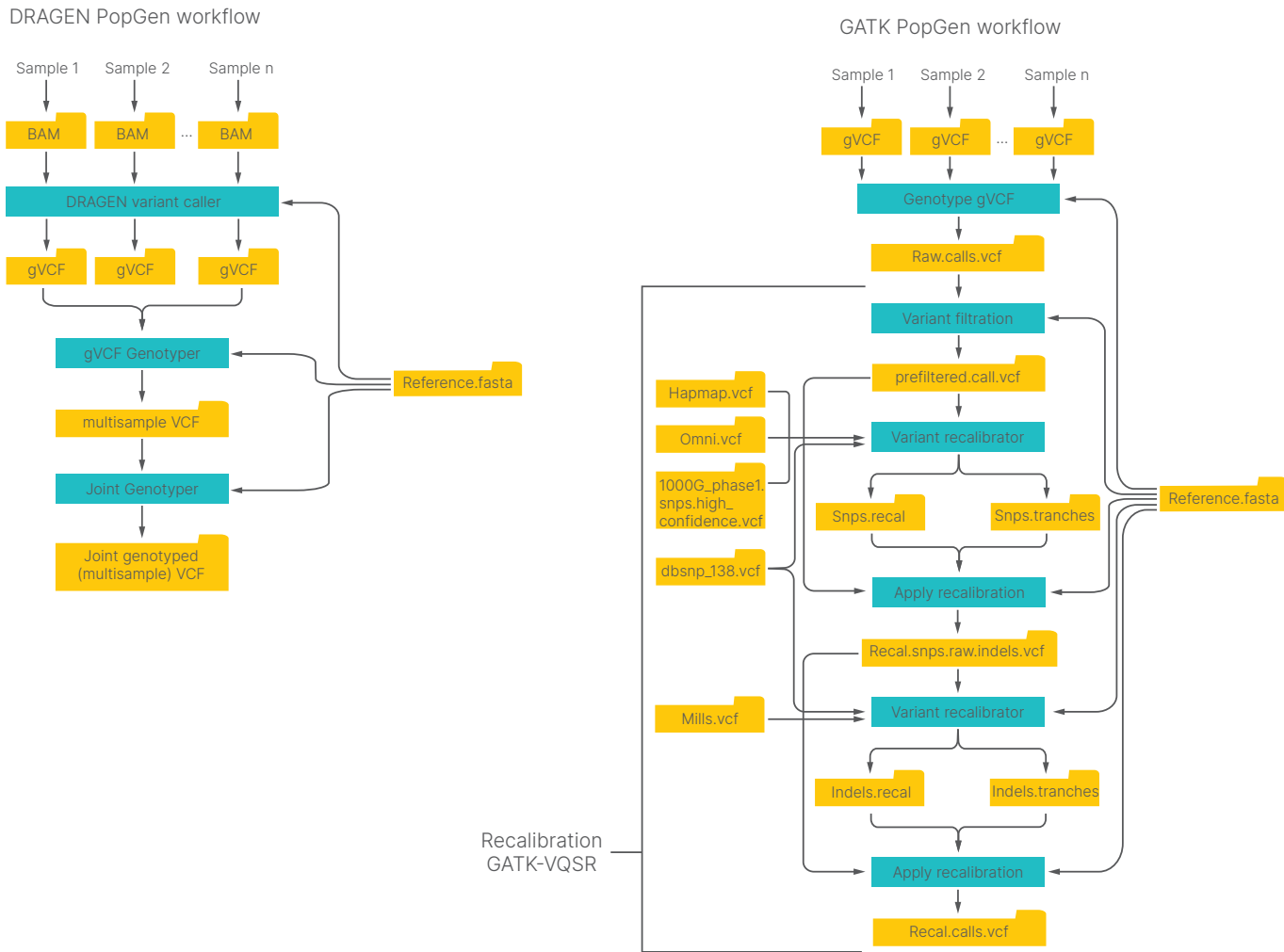



Figure 1: PopGen data processing and analysis workflows using the DRAGEN Platform (left) and GATK best practices (right) workflows.³— The DRAGEN PopGen workflow is composed of two distinct steps: aggregation of gVCFs through the cohort with gVCF Genotyper (DRAGEN-GG), and the joint genotyping step with Joint Genotyper (DRAGEN-JG). The DRAGEN workflow does not proceed with any recalibration steps.

High-coverage WGS

To demonstrate the performance of the DRAGEN Platform in high-coverage WGS samples, we performed a direct accuracy comparison between the DRAGEN Platform and GATK call sets. Performance was measured using receiver operator characteristic (ROC) metrics in a well-characterized sample (NA12878), with truth variants released by the GIAB, that was part of the original cohort. We restricted analyses to chromosome 17 to minimize computational costs.

 ROC curves plot the true positive rate against the false positive rate at various thresholds. The area under an ROC curve is a metric for variant-calling accuracy.

Results

Four population data sets were assessed by extracting the column containing the truth sample NA12878 from the multisample VCF output and plotting ROC curves. Two were from the GATK workflow and two from the DRAGEN Platform:

- All pass variants from Joint Genotyping (GATK-JG)*
- All pass variants from Joint Genotyping that also passed recalibration only (GATK-VQSR)
- All pass variants from gVCF Genotyper (DRAGEN-GG)
- All pass variants after Joint Genotyping (DRAGEN-JG)

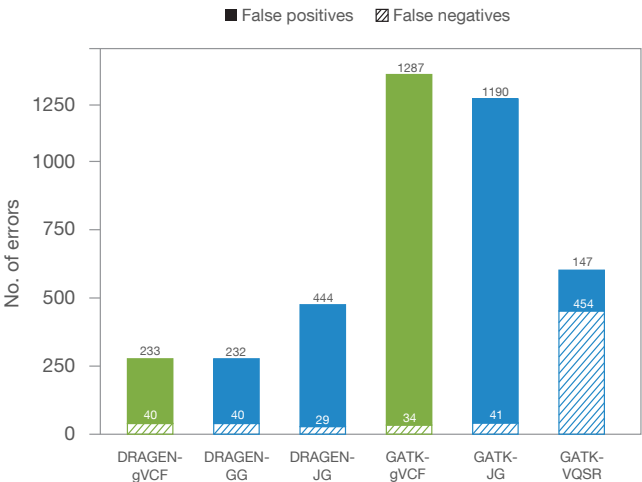
Overall, the DRAGEN Platform outperforms GATK regardless of the workflow composition, driven by superior accuracy in single-sample variant calling for SNPs (Figure 2A) and indels (Figure 2B). An unexpected observation is that DRAGEN accuracy is reduced after Joint Genotyping, due to an increase in false positives (Figure 2 and Figure 3). Traditional joint calling methods available today do not provide any gains when applied to DRAGEN single-sample gVCFs and result in unnecessary higher costs. This is because the DRAGEN platform genotyper includes models of PCR-induced errors and pileup correlated errors.

 Read the [Accuracy improvements in germline small variant calling with the DRAGEN platform application note](#)

* The output of GATK aggregation, before joint genotyping, was not available.

Mendelian errors in trios are a useful metric for broad assessment of precision because they are not restricted to variants within high-confidence regions of the genome. Evaluating the number of Mendelian errors over the total number of sites that are variant in at least one member of the trio in the cohort⁶ is consistent with previous data. Regardless of the workflow, accuracy is increased with the DRAGEN Platform, but performance is reduced after Joint Genotyping (Table 1).

A. SNP errors



B. Indel errors



Figure 2: Variant-calling accuracy in high-coverage WGS data set—False positives and negatives for variant calling of (A) SNPs and (B) indels in a single sample gVCF (green bars) and multisample VCF (blue bars) after PopGen processing with the DRAGEN Platform (GG, JG) and the GATK workflow (JG, VQSR).

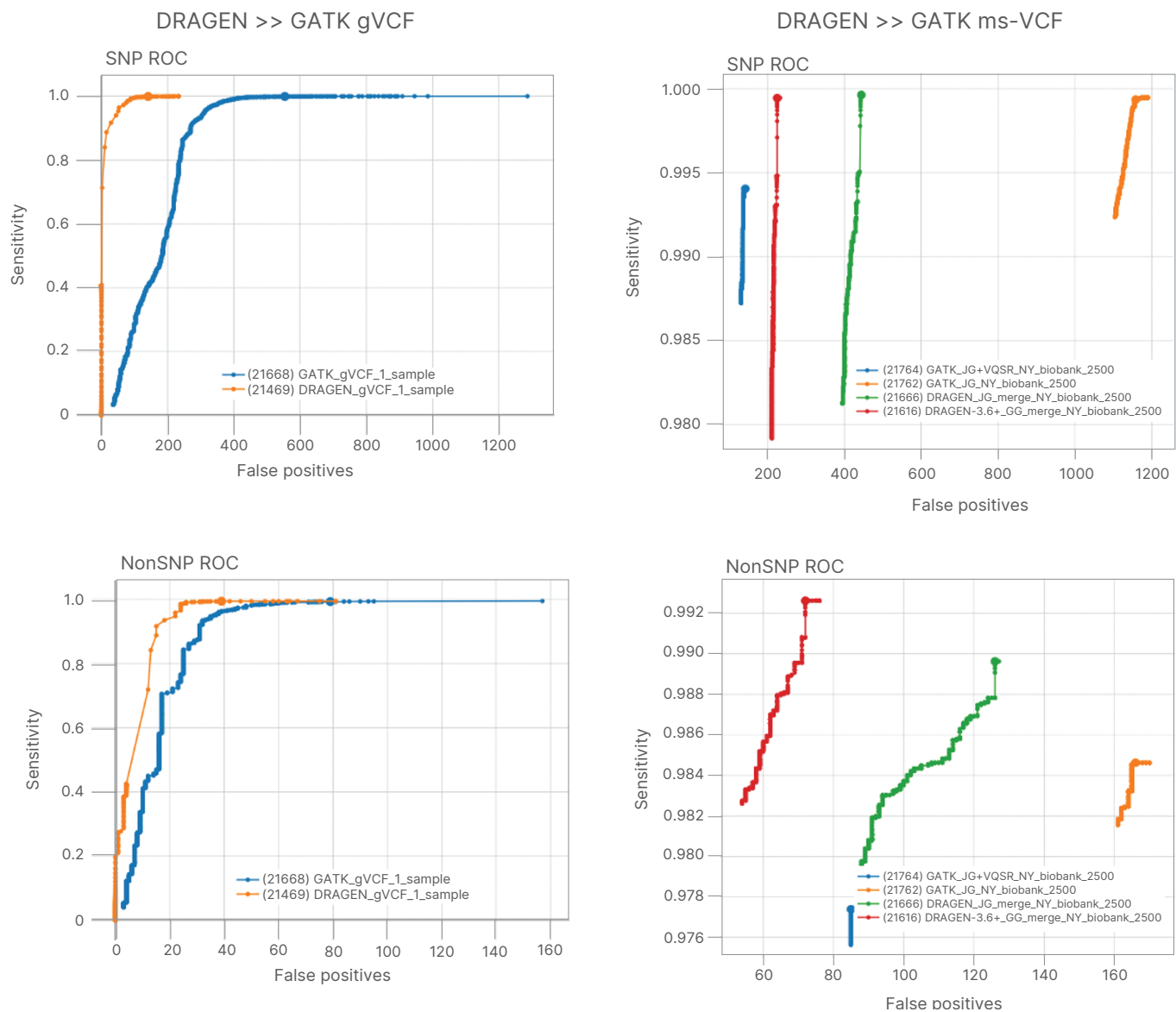


Figure 3: ROC curves after cohort analysis in high-coverage WGS—Computed ROC curves for single-sample gVCFs (left panels) and multisample VCFs (right panels) output from cohort analysis workflows.

Table 1: Calculation of Mendelian errors in a trio present in the high-coverage WGS cohort

Mendelian errors	GATK Joint Genotyper	GATK VQSR	DRAGEN gVCF Genotyper	DRAGEN Joint Genotyper
Inside confident regions	1808/139,375 (1.30%)	833/133,195 (0.63%)	315/127,220 (0.25%)	385/127,667 (0.30%)
Whole chromosome 17	10,433/220,814 (4.72%)	5272/184,275 (2.86%)	4540/179,197 (2.53%)	5318/186,933 (2.84%)

Effect of sample size on cohort analysis

The effect of sample size on the performance of DRAGEN Joint Genotyping was evaluated by comparing genome-wide accuracy metrics with increasing numbers of 3, 6, 10, 50, and 100 samples. Compared to baseline metrics with a single sample, we saw a decrease in false negatives and an increase in false positives for SNPs (Figure 4A), and increases in both metrics for indels (Figure 4B). As before, joint calling methods do not provide gains for DRAGEN single-sample gVCFs, which include models of PCR-induced and pileup correlated errors.

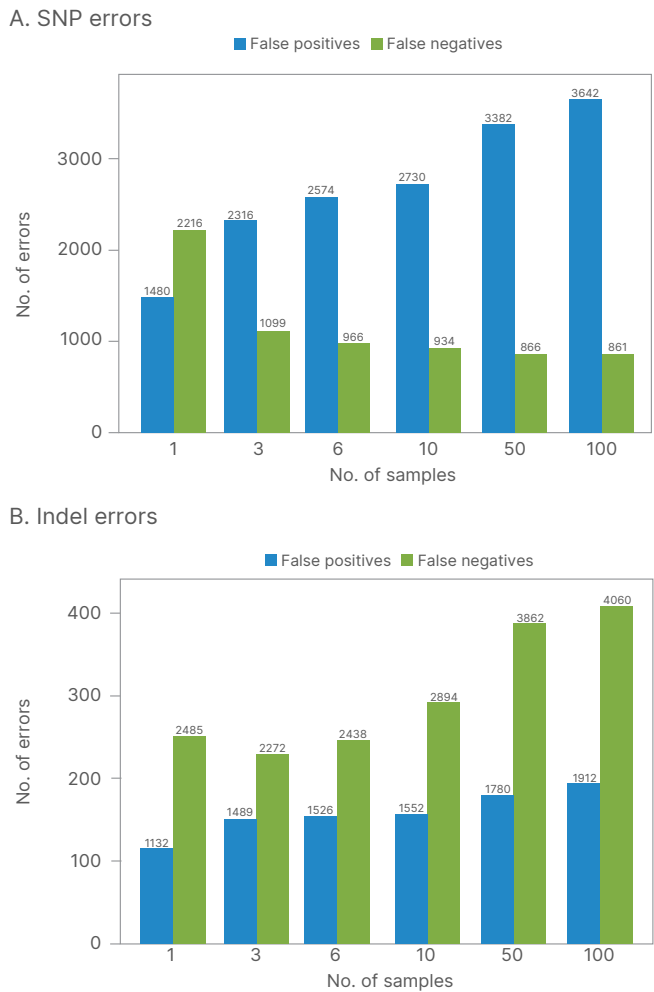


Figure 4: Effect of sample size on joint genotyping—False positives and negatives for (A) SNPs and (B) indels plotted after joint genotyping with the DRAGEN Platform for increasing sample sizes in high-coverage WGS data set.

Low-coverage WGS

To investigate the potential benefit of joint genotyping at lower coverages, we downsampled the alignments from the 1000 Genomes cohort to 15× and reprocessed them with the DRAGEN Platform. A region consisting of the first 10 Mbp of chromosome 17 was selected for this analysis. gVCFs from the downsampled data were aggregated and joint genotyped and performance was measured using ROC metrics for the NA12878 truth sample.

Results

The performance in a low-coverage WGS data set was measured by extracting the column containing NA12878 truth sample from the multisample VCF and plotting error counts after both gVCF Genotyper and Joint Genotyper. Results are similar to high-coverage data, with gains in SNP sensitivity outweighed by losses in specificity (Figure 5A) and indel calling showing regressions on all metrics (Figure 5B).

High-coverage WES data set

The performance of the DRAGEN Joint Genotyper in WES data was measured using a panel of 10 samples comprising eight unrelated samples from the CEU population and two children from the GIAB trios. Joint genotyping was performed on subsets comprising of 1, 3, 4, 6, 8, and 10 samples. Performance was measured within the exome capture regions, using ROC metrics in the NA12878 truth sample.

Results

Calls from the different subsets were assessed by extracting the column containing the truth sample NA12878 from the multisample VCF output and plotting ROC curves. As in the other analyses, no visible benefit from joint genotyping more samples was observed (Figure 6). The preferred DRAGEN PopGen workflow stops after running the gVCF Genotyper and omits the joint genotyping step (Figure 7).

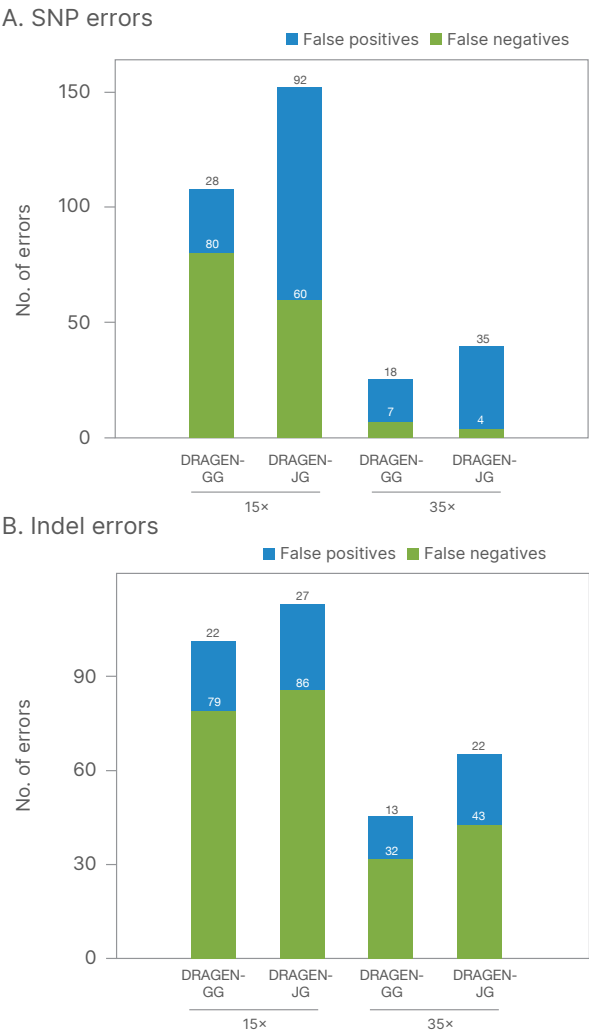


Figure 5: Variant calling accuracy in low-coverage WGS data set—False positives and negatives for variant calling of (A) SNPs and (B) indels in multisample VCF after PopGen processing with the DRAGEN Platform (GG, JG) comparing sequencing coverages of 15x and 35x.

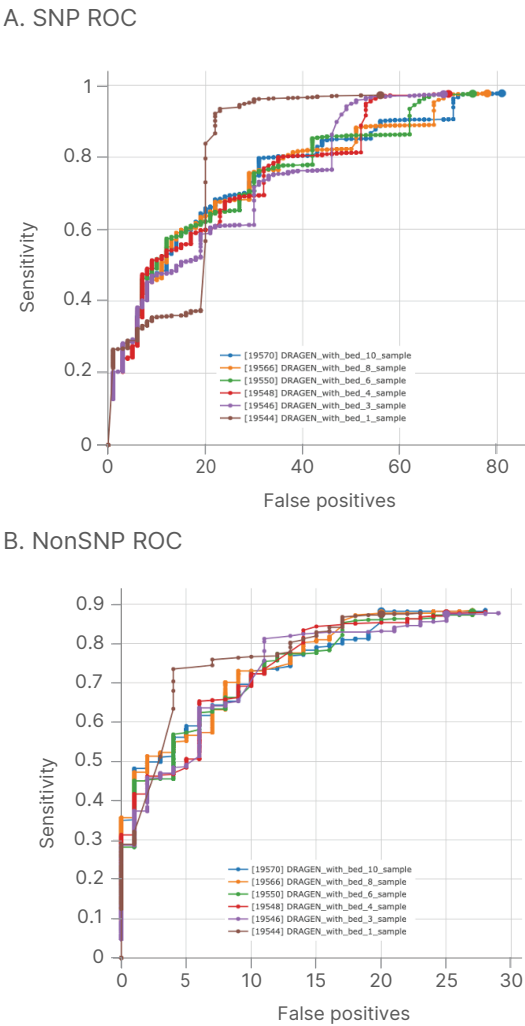


Figure 6: Effect of DRAGEN Joint Genotyper on high-coverage WES—ROC curves for increasing number of samples after joint genotyping with the DRAGEN Platform.

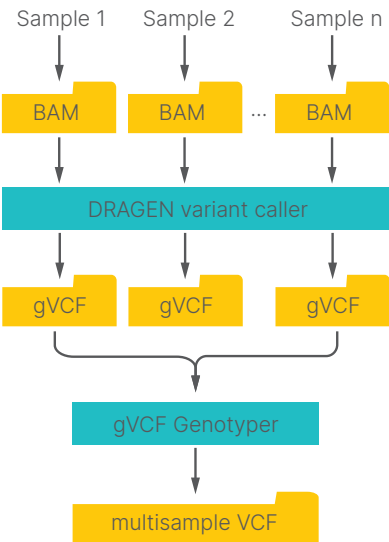


Figure 7: Recommended DRAGEN PopGen workflow

Summary

The established GATK Best Practices workflow for cohort data processing and analysis includes a joint genotyping step, where cohort information is used to improve genotype calls in individual samples. However, based on the results presented in this technical note, joint genotyping as implemented by the GATK workflow is not recommended for use with the DRAGEN Platform for large cohorts of well-covered samples (at least 30× coverage), due to risks of introducing errors, high computation times, and costs. The preferred DRAGEN PopGen workflow stops after running the gVCF Genotyper and omits the joint genotyping step (Figure 7). This results in aggregation of individual gVCFs and produces a multisample VCF with analysis-ready variants. This simplified workflow with the DRAGEN Platform delivers highly accurate population call sets in a flexible and efficient manner.

illumina®

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

© 2022 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-00561 v1.0

References

1. The 1000 Genomes Project Consortium; Auton A, Brooks LD, et al.. [A global reference for human genetic variation](#). *Nature*. 2015;526:68–74. doi: 10.1038/nature15393.
2. The 1000 Genomes Project Consortium. [A map of human genome variation from population-scale sequencing](#). *Nature*. 2010;467:1061–73. doi: 10.1038/nature09534.
3. Intel, 2016. Infrastructure for Deploying GATK Best Practices Pipeline. intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf. Accessed December 01, 2020.
4. DePristo MA, Banks E, Poplin R, et al. [A framework for variation discovery and genotyping using next-generation DNA sequencing data](#). *Nat Genet*. 2011;43:491–501.
5. Zook JM, McDaniel J, Olson ND, et al. [An open resource for accurately benchmarking small variant and reference calls](#). *Nat Biotechnol*. 2019;37:561–6. doi: 10.1038/s41587-019-0074-6.
6. Roslin NM, Welli L, Paterson AD, Strug LJ. [Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes](#). *bioRxiv*. 2016;078600–078600. doi: <https://doi.org/10.1101.078600>.