Sentieon DNAscope: high accuracy small variant calling using machine learning supporting short read platforms

Brendan Gallagher, Donald Freed, Zhipan Li
Sentieon Inc., San Jose, CA 95134

Introduction

We present DNAscope[1], an accurate and efficient short-read germline small-variant caller. DNAscope combines the robust and well-established preprocessing and assembly mathematics of the GATK's HaplotypeCaller with a machine-learned genotyping model. Benchmarks of DNAscope and DNAseq (Sentieon's GATK-matching germline variant calling pipeline) demonstrate that DNAscope achieves superior SNP and insertion/deletion accuracy with reduced computational cost. DNAscope's machine-learning genotyping model is tuned for established and newly developed short read sequencing platforms including Illumina, Element Biosciences, and Ultima Genomics.

Workflow

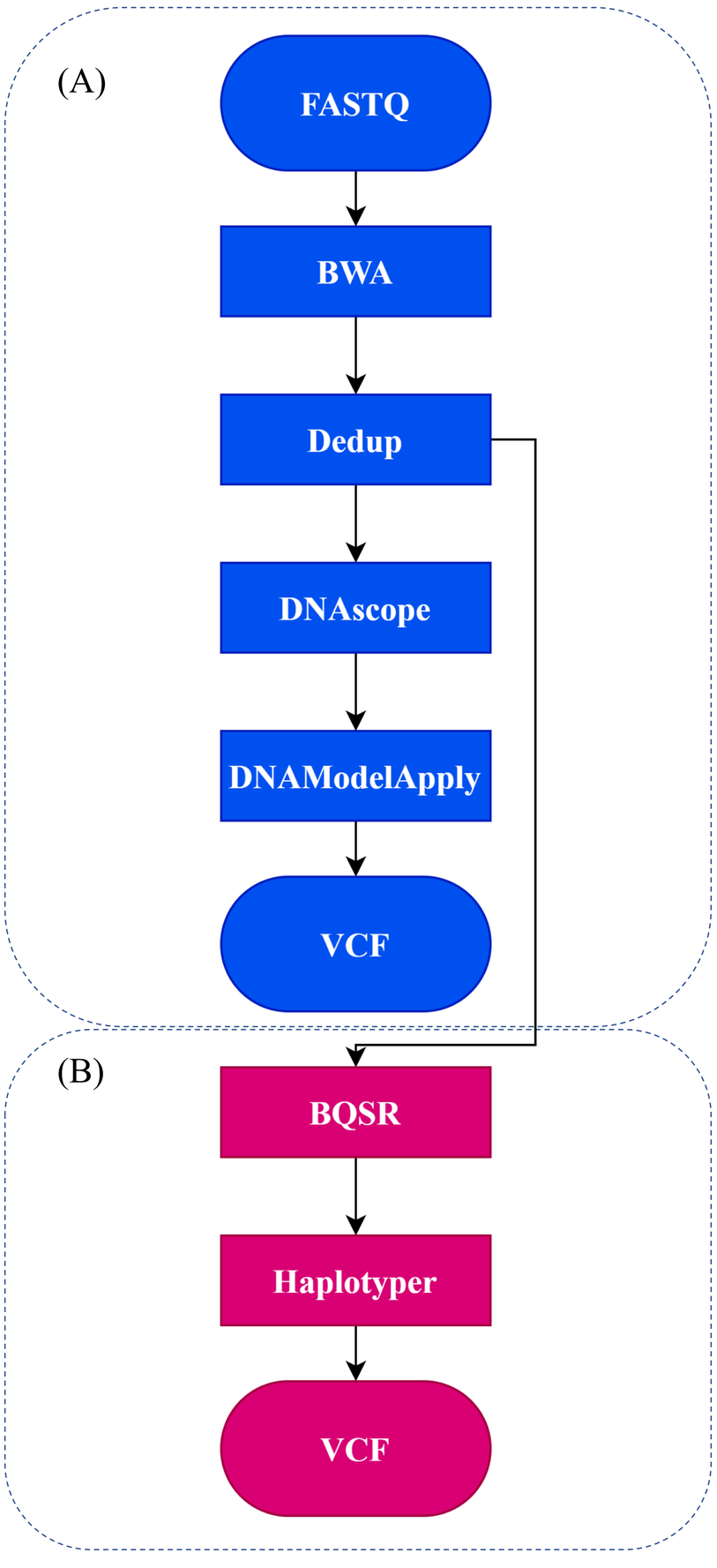

Figure 1: Overview of the benchmarked workflows. (A) The Sentieon DNAscope pipeline. (B) The Sentieon DNAseq pipeline. Sentieon's DNAseq pipeline provides identical results to the GATK best practices pipeline for germline variant calling, while DNAscope pipeline provides significantly higher variant calling accuracy. A sequencing platform dependent model is used for DNAscope pipeline.

Improved Variant Calling Accuracy

The variant calling accuracy as measured against the GIAB v4.2.1 truthset for the HG002 sample is shown in Figures 2 and 3. Sentieon's DNAscope provides highly accurate SNP and indel variant calling across all platforms. Sentieon's DNAseq pipeline provides identical results to the GATK best practices pipeline for germline variant calling but has lower accuracy on the GIAB benchmark dataset.

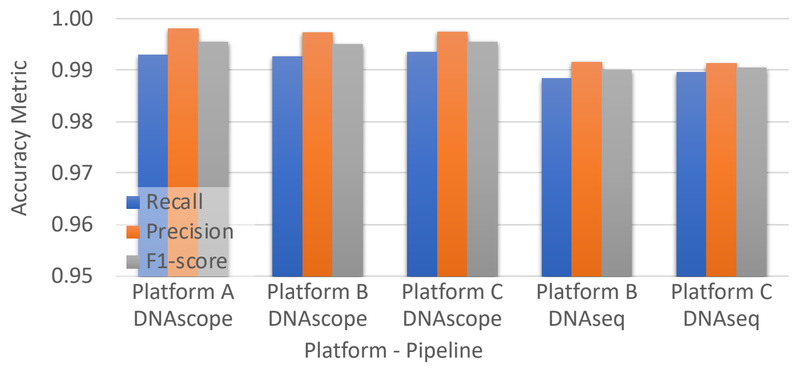

Figure 2: SNP calling accuracy. Variant calling accuracy of the Sentieon DNAseq and DNAscope pipelines for SNVs, measured with the GIAB v4.2.1 benchmark VCF and BED files for HG002. Sentieon DNAseq matches the GATK HaplotypeCaller but has lower accuracy on the Genome in a Bottle benchmark.

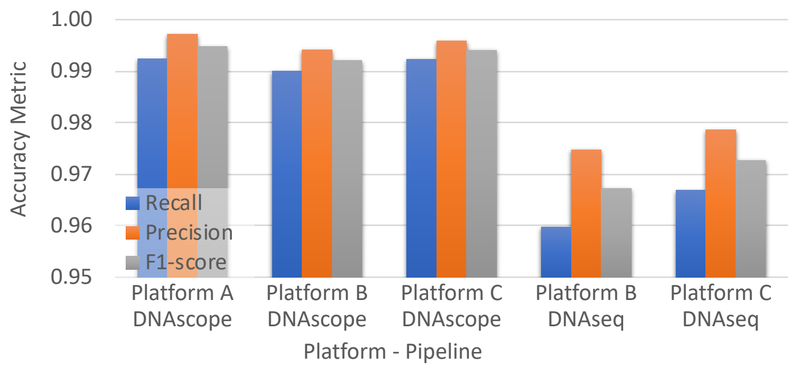

Figure 3: INDEL calling accuracy. Variant calling accuracy of the Sentieon DNAseq and DNAscope pipelines for INDELs, measured with the GIAB v4.2.1 benchmark VCF and BED files for HG002. Sentieon DNAseq matches the GATK HaplotypeCaller but has lower accuracy on the Genome in a Bottle benchmark.

DNAscope Methodology

The Sentieon DNAscope uniquely combines the well-established methods from haplotype-based variant callers with machine learning to achieve improved accuracy.

- DNAscope uses the same overall architecture as GATK's HaplotypeCaller, including active region detection, local assembly, calculation of read-haplotype likelihoods.
- DNAscope improves robustness through removing down-sampling, an improved implementation and local assembly method, better resource management and improved computational algorithms.
- DNAscope's genotype model is designed to help distinguish systematic noise from true germline variants. Gradient Boosting Machines (GBMs) are used to learn error patterns from library prep, sequencing, and alignment as well as to extract informative signals from regions that are inherently difficult to interpret.

Accuracy with Novel Sequencing Chemistry

Platform D generates an impressive amount of data at a highly competitive price point using a novel sequencing chemistry. In their initial publication, Platform D has defined a high-confidence region excluding long homopolymers and other difficult regions[2]. The accuracy of DNAscope pipeline across this truth set is shown in Figure 4, along with their customized GATK.

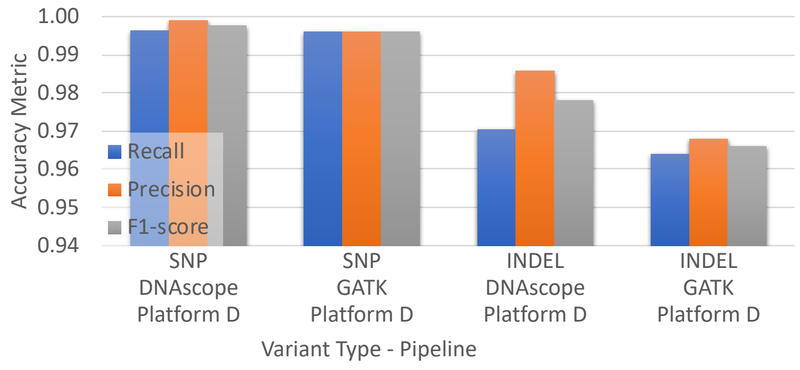

Figure 4: Variant calling accuracy of the Sentieon DNAscope pipeline with the Platform D. Calling accuracy of both SNPs and INDELs is measured with the GIAB v4.2.1 benchmark VCF file and the platform-specific high-confidence region BED file for HG002.

Runtime and Computing Costs

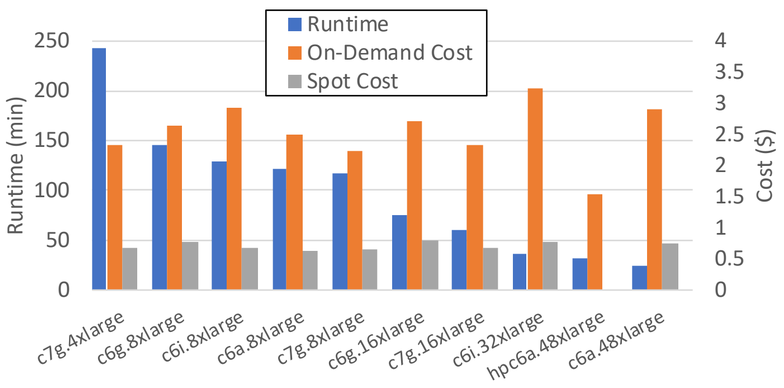

Figure 5: Runtime benchmark of the Sentieon DNAseq pipeline for FASTQ to VCF. Instances are sorted by overall runtime for the Sentieon DNAseq pipeline. On-Demand compute costs and spot compute costs are relatively consistent across all instances tested.

Conclusions

- DNAscope combines the well-established mathematical and statistical models used in the GATK's HaplotypeCaller with machine learning for variant genotyping to achieve superior accuracy while maintaining computational efficiency.
- Sentieon DNAscope pipeline provides robust and efficient variant calling with state-of-the-art accuracy across a variety of sequencing platforms.
- The pipelines are highly scalable and can be used on a variety of instance types. The software can scale up to the 192 vCPU c6a.48xlarge instances for turnaround times of under 24 minutes or down to c7g.4xlarge instances for more flexibility on the spot market.

References

[1] Freed, D. *et al*. bioRxiv. 10.1101/2022.05.20.492556 (2022)
[2] Almogy, G. *et al*. bioRxiv. 10.1101/2022.05.29.493900 (2022)

*Brendan Gallagher – brendan.gallagher@sentieon.com