Rapid Whole Genome and Whole Exome Variant Detection Using a Novel Fluorescently Labeled Reversible Terminated Nucleotide Sequencing System and GPU-based Accelerated Analysis

Kenneth Gouin III[1], Ann Tong[1], Ankit Sethia[2], Ryan Shultzaberger[1], Mehrzad Samadi[2], Tong Zhu[2], Martin Fabani[1], Timothy Looney[1], Eli Glezer[1]
[1]Singular Genomics Systems Inc., San Diego, California. [2]NVIDIA, San Jose, California.

Abstract

There remains a need to reduce next-generation sequencing (NGS) turnaround for time sensitive applications (1,2). Reducing turnaround requires faster sequencing and accelerated data analysis. We recently introduced the Singular Genomics G4™ platform for rapid sequencing-by-synthesis (SBS), which can deliver four human whole genomes at ~30x coverage in under 19 hours. Here we present accelerated pipelines for whole genome and whole exome germline variant detection on the G4 that leverage the NVIDIA Clara Parabricks platform and custom DeepVariant (3) models.

Methods

We generated PCR-free whole genome (KAPA biosciences) or whole exome (IDT xGen exome kit) sequencing libraries from Covaris-sheared or enzymatically fragmented gDNA, respectively. WGS libraries from GIAB samples HG001 and HG002 and WES libraries from HG001-HG004 were sequenced via 2x150bp reads on the G4 platform to achieve ~31x and ~110x coverage for WGS and WES libraries, respectively.

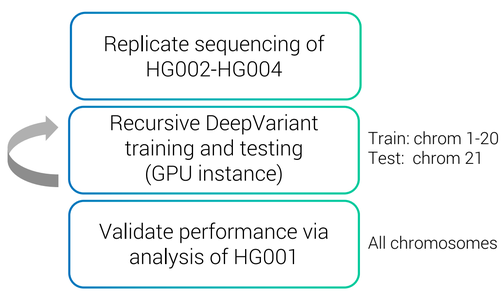
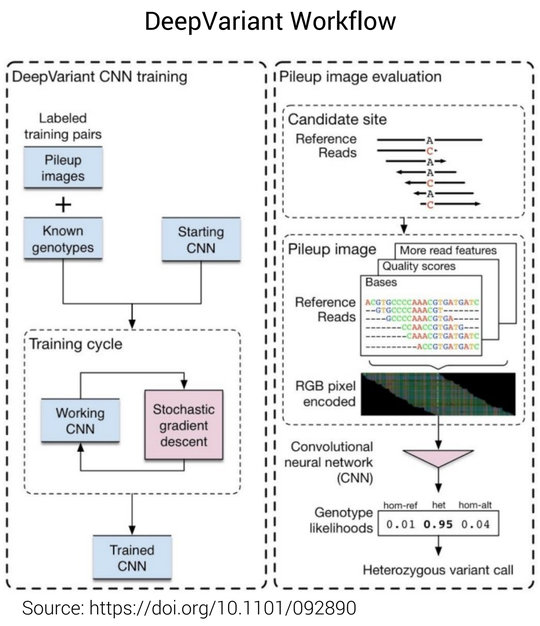


Figure 1. Workflow for training of custom DeepVariant models. Data from either HG001 (WGS) or HG002-4 (WES) were used for recursive model training using a warm start from the corresponding Illumina models. Finally, model performance was assessed on HG002 (WGS) or HG001 (WES).



Source: https://doi.org/10.1101/092890

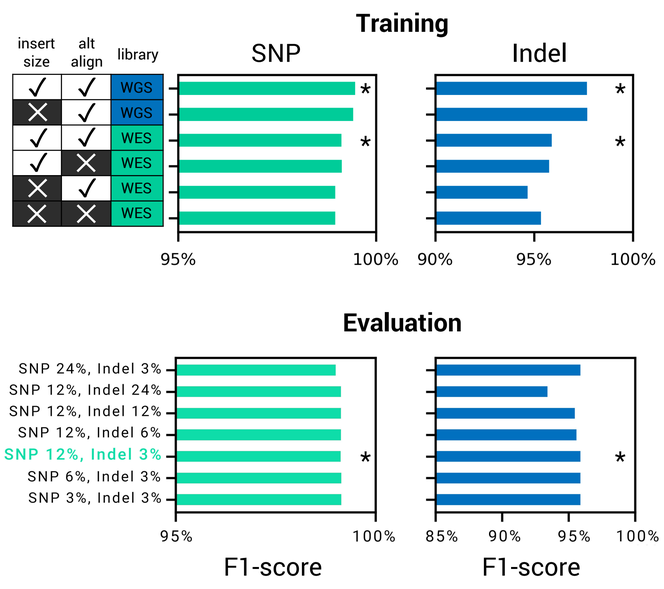


Figure 2. Training and evaluation parameter influence on performance. We assessed the impact of including insert size and alternate alignment as channels for DeepVariant training. Consistent with observations from Illumina datasets, the insert size channel, introduced in DeepVariant v1.4, improves model performance. Inclusion of alternate alignment further improves indel performance. During evaluation, the minimum indel fraction has the greatest impact on performance. To accelerate exploration of the parameter space, we leveraged GPU resources on AWS, orchestrated via a custom Nextflow pipeline. *indicates parameters chosen for final assessment.

Results -- Performance of Default and Custom DeepVariant Models

To test performance, whole exome or whole genome libraries were prepared from enzymatically-fragmented HG001 or Covaris-sheared HG002 gDNA, respectively, followed by sequencing via the F2 flowcell (150M reads) with 2x150bp reads. Alignment was performed with Parabricks, trained DeepVariant models were used for variant detection, and performance was assessed with hap.py.

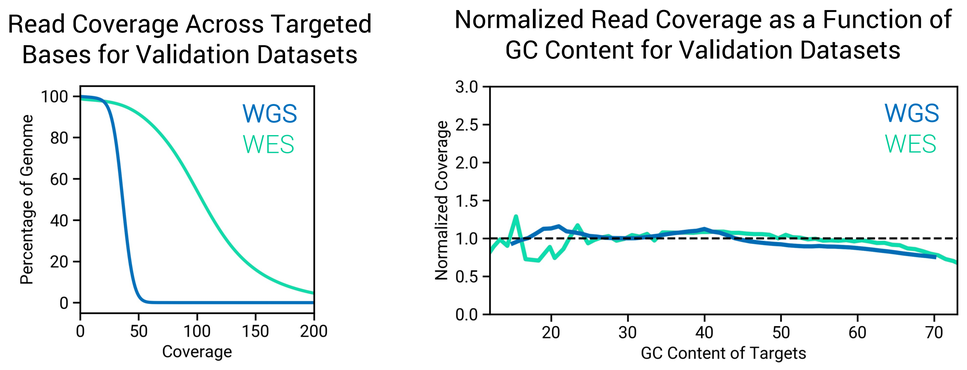


Figure 3. Read coverage for the 31x whole genome and 110x whole exome datasets used for validation.

Metric	31x WGS, Illumina Model	31x WGS, Singular Model	110x WES, Illumina Model	110x WES, Singular Model
SNP Precision	99.86%	99.86%	99.48%	99.67%
SNP Recall	99.12%	99.10%	98.65%	98.60%
SNP F1-Score	99.49%	99.48%	99.06%	99.13%
Indel (<50bp) Precision	98.37%	98.56%	98.61%	98.87%
Indel (<50bp) Recall	96.27%	96.81%	93.62%	93.09%
Indel (<50bp) F1-Score	97.31%	97.68%	96.05%	95.89%

Table 1. hap.py metrics for default and custom DeepVariant models, as assessed on high confidence regions of HG002 (WGS) or HG001 (WES).

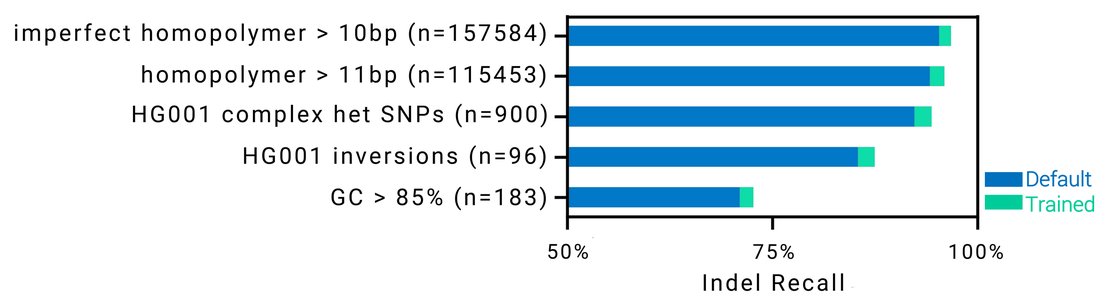


Figure 4. Performance of default and trained WGS models on select challenging genome features (4). The G4 shows robust performance with homopolymers. Model training preferentially improved Indel performance, particularly over homopolymeric and high GC content regions.

Conclusion

We have successfully trained and implemented accelerated DeepVariant v1.4 whole genome and whole exome models for the G4 Sequencing Platform, resulting in improved SNP and indel precision and a rapid fastq-to-vcf turnaround of 29 and 4 minutes for 30x whole genome and 100x whole exome analysis, respectively, using an 8 GPU AWS p4d.24xlarge instance. To further improve performance, efforts are underway to build and test de novo DeepVariant models for WGS and WES. We anticipate that the combination of rapid-SBS and GPU-based acceleration will significantly reduce turnaround for the most time sensitive NGS applications.

References

1. Bamshad et al. Nat Rev Gen (2011) doi: 10.1038/nrg3031
2. Xu. CSBJ. doi:10.1016/j.csbj.2018.01.003
3. Poplin et al. BioRxiv (2018) doi: 10.1101/092890
4. GA4GH GIAB Stratification Regions: https://github.com/genome-in-a-bottle/genome-stratifications

Special thanks to Andrew Carroll (Google AI) for advice on training and testing of DeepVariant.