SINGULAR GENOMICS

Training and validation of an accelerated DeepVariant model for germline exome analysis on the G4™

Kenneth Gouin III[1], Ann Tong[1], Sabrina Shore[1], Ryan Shultzaberger[1], Martin Fabani[1], Eli N Glezer[1], Timothy Looney[1]
(1) Singular Genomics Systems, Inc., 3010 Science Park Rd, San Diego, CA 92121

Background

Exome sequencing to identify germline variation is a key tool for the diagnosis of genetic disease, population genome studies (1) , and as a component of tumor-normal sequencing protocols used in precision oncology (2). Traditional variant detection methods rely upon manually tuned, parameterized statistical models to achieve high accuracy. Recently, this paradigm has been challenged by DeepVariant, a method leveraging deep convolutional neural networks trained upon read pileup images to identify variants (3). DeepVariant models have been trained to achieve high accuracy with diverse sequence data types. Here we present a highly performant DeepVariant model optimized for exome analysis on the novel G4 Sequencing Platform.

Methods – Sequencing and DeepVariant model training

Figure 1. Exome sequencing workflow on the G4. Multiple exome capture kits were used for exome library preparation of Genome in a Bottle (GIAB) samples HG001-6, followed by sequencing on the G4 to >100x coverage via 2x150bp reads.

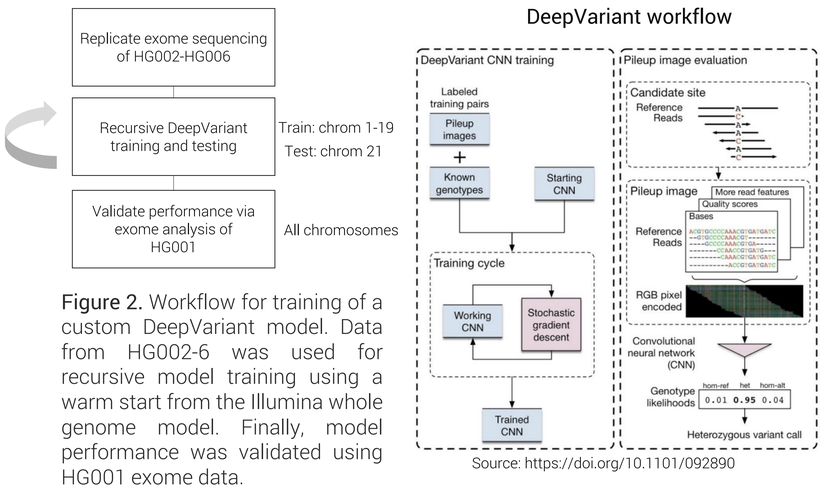

Figure 2. Workflow for training of a custom DeepVariant model. Data from HG002-6 was used for recursive model training using a warm start from the Illumina whole genome model. Finally, model performance was validated using HG001 exome data.

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

Results – Sequencing quality metrics and coverage

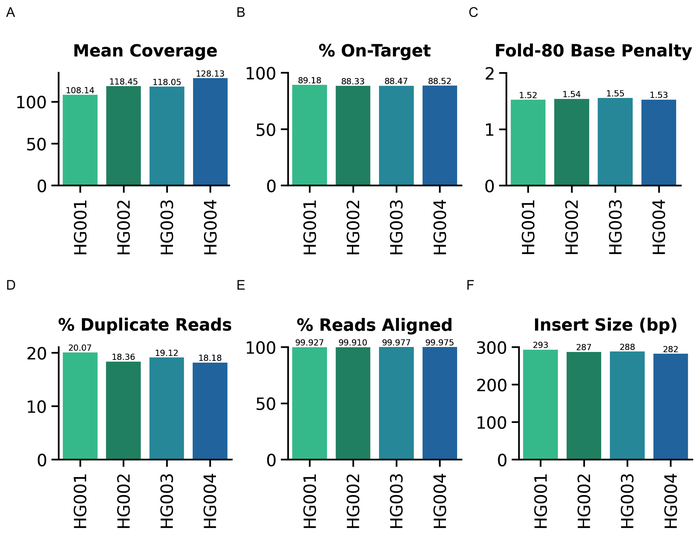

Figure 3. Sequencing quality metrics. Four exome libraries were prepared for GIAB samples HG001-HG004 using the IDT xGen exome kit, followed by 2x150bp sequencing via the F2 flow cell (150M reads). Picard tools was used to determine the mean target coverage, percent on-target reads, fold-80 base penalty, percent duplicate reads, percent aligned reads, and mean insert size distribution for each library (A-F, respectively). HG001 data was used to validate performance. All data met system quality specifications: 88 and 80% Q30 for R1 and R2, respectively; accuracy >99.7%.

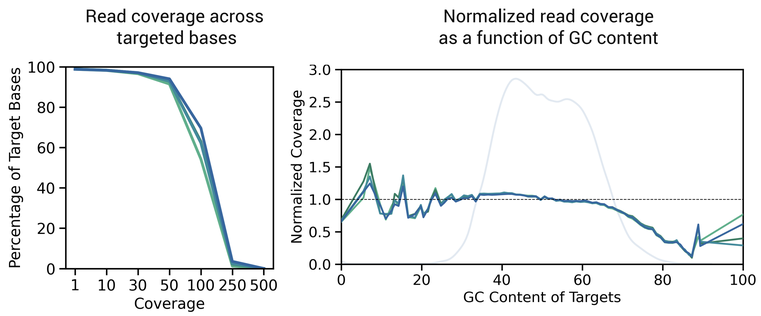

Figure 4. Coverage and GC bias metrics. (A) Read coverage across targeted bases. Coverage values are derived from Picard tools CollectHSMetrics. (B) Coverage uniformity as a function of GC content. Values represent the relative read coverage over panel target regions of a given GC content, normalized to the mean coverage across all target regions. Gray line indicates frequency of targets by GC content.

Results – Variant detection performance for HG001

Metric	50x mean target coverage	100x mean target coverage
%Bases ≥ 10x coverage	97.40%	98.14%
SNP Precision	99.39%	99.53%
SNP Recall	98.30%	98.53%
SNP F1-Score	**98.84%**	**99.03%**
Indel (<50bp) Precision	97.45%	97.76%
Indel (<50bp) Recall	91.22%	93.09%
Indel F1-Score	**94.23%**	**95.36%**
Total SNPs	22411	22493
Het:Hom Ratio	1.59	1.58
Ti:Tv Ratio	3.01	3.00

Table 1. Variant detection metrics. Germline variant detection metrics for HG001. An exome library was prepared for GIAB sample HG001 using the IDT xGen exome kit, followed by 2x150bp sequencing via the F2 flow cell. Reads were aligned to GRCh38 with BWA and subsequently downsampled to 50x and 100x mean target coverage followed by variant detection using the trained DeepVariant model, implemented on the Parabricks platform (~8min fastq to vcf turnaround using 4 GPUs). Performance was assessed using hap.py with the NIST GIAB v4.2.1 truth set.

Conclusions

We have produced a highly performant custom DeepVariant model for exome analysis on the G4. The model demonstrates high accuracy for both SNP and indel calling with the gold standard HG001 reference, meeting or exceeding the performance of custom DeepVariant models produced for other sequencing platforms (3). In order to minimize the possibility of overfitting, training was performed using HG002-6 data, with HG001 reserved exclusively for validation.

Exome analysis is sensitive to biases in the target enrichment process but also sequencing errors associated with certain nucleotide motifs, particularly those that lead to uneven coverage. In this context the strong variant detection performance reflects the compatibility of the G4 platform with common exome library preparation kits and the suitability of the sequence data for variant detection applications.

References and Acknowledgements

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. Kumaran et al. BMC Bioinf (2019) doi: 10.1186/s12859-019-2928-9

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