Rapid Targeted Germline and Somatic Variant Detection Using the G4™ Sequencing Platform

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]
[1]Singular Genomics Systems Inc., San Diego, California. [2]NVIDIA, San Jose, California.

Background

Next generation sequencing (NGS) has become an indispensable tool for the diagnosis of genetic disease, though there remains a need to reduce turnaround for time sensitive applications. Reducing turnaround requires faster sequencing and accelerated data analysis. We recently introduced the Singular Genomics G4™ Sequencing 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 exome germline and targeted somatic variant detection on the G4 that leverage the NVIDIA Clara Parabricks platform.

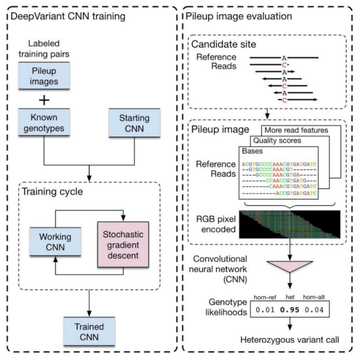
Methods – DeepVariant Training and Implementation for WES

DeepVariant (Poplin, 2018) has emerged as a leading method for germline variant detection. To rapidly build and evaluate candidate custom DeepVariant models we created an automated NextFlow-based training pipeline leveraging GPU resources in the AWS cloud. Custom models were produced for DeepVariant v1.3 and v1.4 using exome sequencing data from HG002-4, then validated on HG001 exome sequencing data (IDT xGen exome; 2x150bp reads). Finally, the validated model was deployed on the Parabricks platform.

Singular Internal Workflow

- Automated construction of training sets from HG002-4 in cloud
- Accelerated training leveraging GPUs in AWS cloud
- Model validation on HG001

DeepVariant Workflow



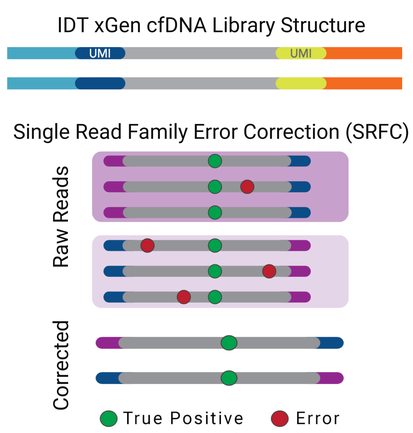
Parabricks External Workflow

- Custom DeepVariant model hosted on NVIDIA Parabricks
- **Fastq-to-vcf:** ~30 min for 30x WGS ~4 min for 100x WES p4d.24xlarge GPU instance

Source: Poplin, 2018

Methods – Somatic Variant Detection via Single Read Family Correction

The Parabricks platform enables rapid somatic variant detection via GPU acceleration of the fgbio single read family consensus (SRFC) duplex sequencing workflow. To test the compatibility of this workflow with the Singular platform, we performed duplex UMI tagging and targeted capture (IDT xGen cfDNA kit with xGen Pan Cancer Panel, 50ng input) of a reference material comprising an equimolar pool of 23 reference cell lines. Libraries were prepared for the G4 and NextSeq 2000, sequenced to ~20,000x coverage via 2x151bp reads, then processed with Parabricks. Finally, variants were detected using varDict (Lai, 2016).



Source: IDT xGen cfDNA handbook

Results – Exome Sequencing and Analysis of HG001

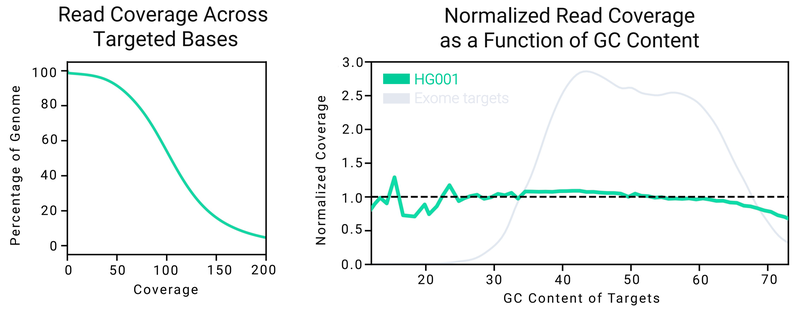


Figure 1. Read coverage across targeted exome regions and coverage as a function of GC content following 2x150bp sequencing of an exome library prepared from enzymatically sheared HG001 gDNA to achieve ~110x coverage.

	DeepVariant v1.3		DeepVariant v1.4	
Metric	100x coverage Default WES model	100x coverage Singular WES model	100x coverage Default WES model	100x coverage Singular WES model
SNP Precision	98.47%	99.54%	99.48%	99.67%
SNP Recall	98.54%	98.41%	98.65%	98.60%
SNP F1-Score	98.50%	98.97%	99.06%	99.13%
Indel (<50bp) Precision	94.91%	97.19%	98.61%	98.87%
Indel (<50bp) Recall	92.82%	92.29%	93.62%	93.09%
Indel F1-Score	93.85%	94.68%	96.05%	95.89%

Table 1. Variant detection performance (assessed via hap.py) for default and custom DeepVariant models produced with DeepVariant v1.3 and v1.4. Training preferentially improved the performance of DeepVariant v1.3 default model, while minimal improvement was achieved over the v1.4 default model.

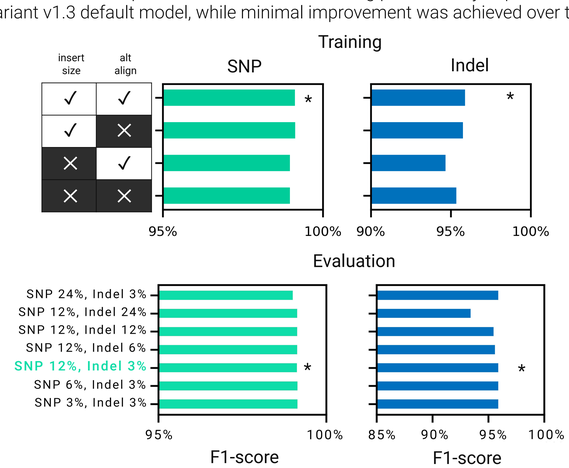


Figure 2. Training and evaluation parameter influence on performance. We assessed the impact of including insert size and alternate alignment as channels for DeepVariant training. Consistent with observations from Illumina datasets, the insert size channel, introduced in DeepVariant v1.4, improves model 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 – Somatic Variant Detection

	Metrics	G4	NextSeq 2000
Picard HS Metrics	Mean Target Coverage	20,072x	20,066x
	% Off Bait	16.4%	14.0%
	% Targets with 0x Coverage	0.26%	0.26%
	% Excluded: Low Base Quality	1.34%	1.28%
	% Excluded: Overlap	37.5%	38.4%
	% Excluded: Off-target	21.0%	19.3%
	Fold 80 Base Penalty	1.48	1.66
	AT Dropout	7.24	11.48
	GC Dropout	0.06	0.01
Variant Metrics	Precision	85.33%	84.34%
	Recall	94.00%	95.66%
	F1-Score	89.45%	89.64%

Table 2. Picard hybrid-selection (HS) metrics and variant calling metrics for libraries sequenced via the G4 and NextSeq 2000 with 2x151bp reads. Single read families with a minimum of 3 supporting reads were retained for variant calling, which was performed with varDict using a minimum allele frequency of 0.005 and a minimum read support of 2.

Results – Observed vs Expected Variant Allele Frequency

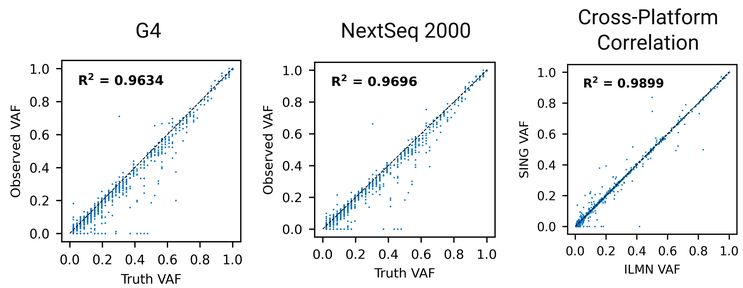


Figure 3. Left and Middle: Observed versus expected variant allele frequencies (VAF) for G4 and NextSeq experiments. Observed VAFs were highly concordant with expected allele frequencies for both systems.
Right: Cross-platform correlation of observed VAFs.

Conclusions

We have successfully implemented a GPU-accelerated DeepVariant whole exome model for the G4. We further demonstrated accelerated single family UMI error correction and somatic variant detection via the Parabricks umi_fgbio workflow. We anticipate that the combination of rapid-SBS and GPU-based acceleration will significantly reduce turnaround for the most time sensitive variant detection applications.

Acknowledgements

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