Rapid Somatic and Germline 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 M Fabani[1], Timothy Looney[1]
[1]Singular Genomics Systems Inc., San Diego, California. [2]NVIDIA, San Jose, California.

Introduction

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. This requires faster sequencing and accelerated data analysis. The G4™ Platform leverages a 4-color rapid sequencing by synthesis (SBS) chemistry and, in combination with advanced optics and fluidics engineering, delivers results of four human whole genomes at ~30x coverage in under 24 hours using the F3 flow cell (up to 450M reads per flow cell). We present accelerated pipelines for whole genome and targeted somatic variant detection on the G4 that leverage the NVIDIA Clara Parabricks platform.

Methods

G4 Sequencing Platform

The G4 Platform is a benchtop sequencer designed to deliver rapid sequencing with throughput flexibility to reduce batching-related delays. The G4 supports single or paired end reads of up to 150 bp, including the ability to include dual index reads for sample multiplexing. Users may analyze up to four flow cells of two types (F2: up to 250M reads, F3: up to 450M reads) in a single run. To facilitate multiplexing, each flow cell comprises four fluidically independent lanes.

Power
up to 480 Gb / Day
up to 3.2 Billion Reads

Speed
< 24 hour run times

Flexibility
1 - 4 flow cells
16 lanes

Accuracy
80 - 90% bases
≥ Q30

	Reagent Configuration[1]	Run Time[2]	Reads / Flow Cell[3]	Reads / Run[3]	Quality[4]
F2 Flow Cell	100 cycles	~11 hours	Up to 250 M	Up to 1,000 M	80-90% Bases ≥ Q30
	200 cycles	~15 hours			
	300 cycles	~19 hours			
F3 Flow Cell	50 cycles	8 - 11 hours	Up to 450 M	Up to 1,800 M	
	100 cycles	11 - 14 hours			
	200 cycles	15 - 19 hours			
	300 cycles	19 - 24 hours			
Max Read[5]	28x91 Single Cell	~ 24 hours	800 M	3,200 M	
	28x50 Spatial FFPE				

[1] Reagents include 50 additional cycles above what is represented to account for adapters and indices.
[2] Run time includes clustering, sequencing and instrument wash for non-indexed reads.
[3] Paired reads passing filter for F2 and F3 are dependent on application and read length.
[4] Performance metrics may be impacted by application, sample quality, library preparation, loading concentration, and other sequencing considerations. Metrics as generated on reference bacterial and human genomes.
[5] Max Read kits specifications are projected, and kits are currently only compatible with 10x Genomics Chromium™ 3 and 5' Gene Expression assays and Visium™ Spatial Gene Expression. Kits allow for 1 sample per lane.

Custom DeepVariant Model Training and Testing

DeepVariant performance may be optimized through the production of custom basecalling models. Custom G4 DeepVariant models were produced for DeepVariant v1.4 using WGS sequencing data from HG001, then validated on HG002 sequencing data (F3 flow cell; 2x150bp reads). Finally, the validated model was deployed on the Parabricks platform.

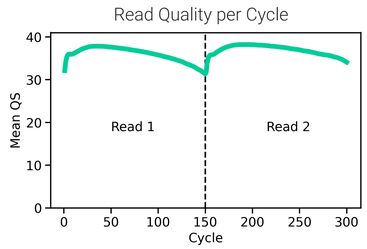
Somatic Variant Detection via Single Read Family Correction

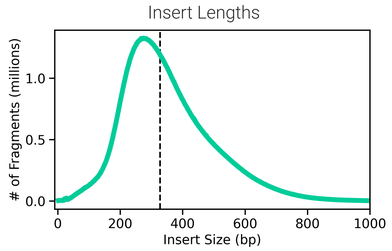
The Parabricks platform enables rapid somatic variant detection via GPU acceleration of the fgbio single read family consensus (SRFC) duplex sequencing workflow. To test the compatibility of this workflow with the G4 Platform, we performed duplex UMI tagging and targeted capture (IDT xGen cfDNA kit with 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).

Results

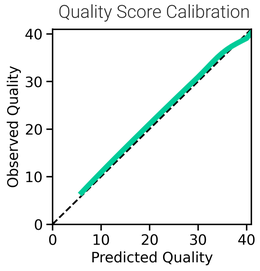
Whole Genome Sequencing with the F3 Flow Cell

Sequencing via a single F3 flow cell with 2x150bp reads format yielded a total of 414M read-pairs, for a mean coverage of 33.6x of the HG002 genome when discounting duplicates (4.6%), ambiguously mapped reads (5.4%), low quality base calls (0.4%), and overlapping bases (7.6%) as reported by Picard.[1] Read quality and accuracy were high (88.6% and 92.6% of base calls ≥ Q30; mean single-pass accuracies of 99.87% and 99.92%, Read 1 and Read 2 respectively**.** Insert lengths were varied, with a median of 328bp. Base quality scores were well calibrated and there was minimal GC related coverage bias**.**

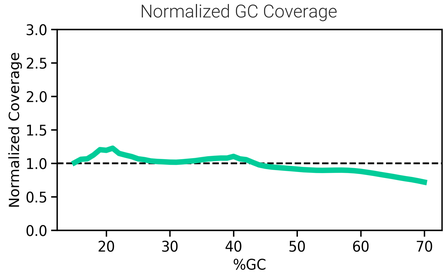




Quality Score Calibration



Normalized GC Coverage



Metric	Value
%PF Reads Aligned	99.9
Duplication Rate (%)	4.55
Median Insert Size (bp)	328
Mean Coverage (X)	33.6
%Bases ≥ 10x Coverage (whole genome)	96.5
%Bases ≥ 10x Coverage (high confidence regions)	99.5
SNP Precision	99.86
SNP Recall	99.18
SNP F1-Score	99.52
Indel (<50bp) Precision	98.33
Indel (<50bp) Recall	97.43
Indel F1-Score	97.88
Total SNPs	3,755,346
Het:Hom Ratio	1.51
Ti:Tv Ratio	2.00

Right: Performance of a custom-trained DeepVariant v1.4 model applied to high confidence regions of HG002 at 30x coverage, implemented using the NVIDIA Parabricks platform. Performance was assessed by hap.py.

Results

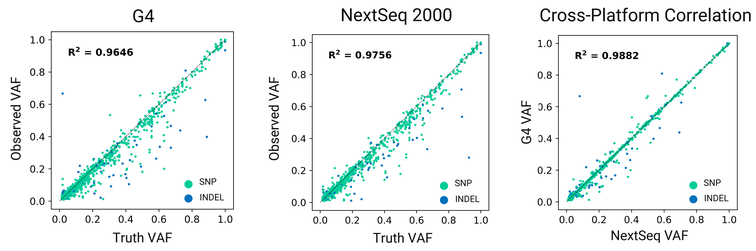
Somatic Variant Detection

	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	79.25%	77.07%
	Recall	90.03%	92.77%
	F1-Score	84.30%	84.19%

Above. Picard hybrid-selection (HS) metrics and variant calling metrics for libraries sequenced via the G4 Platform and NextSeq 2000 with 2x151bp reads. Single read families with a minimum of 3 supporting reads were retained for variant calling, via varDict using a minimum allele frequency of 0.01 and minimum read support of 2.

Results

Observed vs Expected Variant Allele Frequency



Above Left and Middle: Observed versus expected variant allele frequencies (VAF) for G4 Platform and NextSeq 2000 experiments. Observed VAFs were highly concordant with expected allele frequencies for both instruments.
Right: Cross-platform correlation of observed VAFs.

Conclusion

We have successfully implemented a GPU-accelerated DeepVariant whole genome model for the G4 Sequencing Platform. 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 chemistry and GPU-based acceleration will significantly reduce turnaround time for most time-sensitive variant detection applications.

Acknowledgements

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