SINGULAR GENOMICS

APPLICATION NOTE

Whole Exome Sequencing on the G4™

- Rapid SBS enables cost-efficient delivery of 8–64 exome samples in 16-19 hours.
- The G4 Sequencing Platform fits into existing WES and analysis workflows, including accelerated deep learning bioinformatic pipelines.
- G4 delivers highly accurate exome data comparable to leading platforms with SNP and INDEL F1 scores at 99% and 95% respectively.

Introduction

Whole exome sequencing (WES) enables the discovery and assessment of genetic variations linked to rare or complex diseases[1,2] and is a key tool for the diagnosis of genetic disease, population genome studies[3], and tumor-normal sequencing protocols used in precision oncology.[4] Although the exome only represents about 1% of the human genome, mutations in these protein-coding regions are highly associated with disease. New advancements in next generation sequencing (NGS) and artificial intelligence (AI) technologies now allow research beyond clinical diagnosis as WES can also help inform and improve drug discovery, personalized medicine, and reproductive health.[1,3]

The WES workflow consists of 3 major steps: library preparation, sequencing, and data analysis. During library preparation, DNA is isolated and fragmented. Exon fragments are then selected and enriched before sequencing. Analysis of the resulting data requires tools for alignment of sequence reads to the reference genome and for variant calling. Mutations can result in single-nucleotide variants (SNVs), copy number variations (CNVs), and insertion-deletion (indels). Identified variants are then compared to large databases to determine disease-associated variants or pathways.

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 G4 Sequencing Platform.

The G4 Sequencing Platform is a highly versatile benchtop sequencer suitable for demanding research and clinical research applications. The G4 leverages a novel, 4 color rapid SBS chemistry and advanced optical and fluidics engineering to deliver unmatched power and versatility in key applications like whole exome sequencing. The G4 is compatible with existing upstream WES library preparation kits and outputs demultiplexed FASTQ files compatible with existing bioinformatic pipelines, like the NVIDIA Parabricks accelarated Google DeepVariant used in this study.

Exome Parameters

The G4 Sequencing Platform enables users to run 1, 2, 3, or 4 flow cells at a time. Mixing and matching two flow cell types across 4 positions enables users to start the system with 8 different run sizes. Each flow cell has 4 independant lanes, enabling up to 16 independant lanes per run, providing users flexibility in designing sequencing experiments. G4 sequencing output, run time, accuracy, quality, and exome throughput by lane, flow cell, and run are shown in **Table 1**.

Flow Cell Type	F2	F3
Cycles	300	300
Throughput (M Reads)	150-165M per FC 600-660M per run	300-330M per FC 1,200-1,320M per Run
Run Time (Hours)	16-19	16-19
Accuracy	75-90% Bases ≥ Q30	
Quality	99.6-99.9%	
Samples / Lane	2	4
Samples / FC	8	16
Samples / Run[a]	8 - 64	
Samples / Week[b]	8 - 320	

Table 1: Exome sequencing parameters.
[a] Assumes 34 Mb, ~100x coverage.
[b] Assumes 1-5 G4 sequencing runs per week.

Methods

NGS LIBRARY PREPARATION AND SEQUENCING

Using QuantaBio's sparQ DNA Frag & Library Prep Kit, 150ng human genomic DNA HG001-HG004 (NIST) was enzymatically fragmented then inactivated. Singular Genomics Universal Adapters were ligated (0.075uM final adapter concentration) activated, and purified with sparQ PureMag beads. Libraries were amplified with QuantaBio's HiFi PCR Master Mix and Singular GenomicsPCR primers (6 cycles), further purified with sparQ beads and eluted in water.

Exome capture was performed with IDT xGen Hybridization and Wash Kit (250 ng each genomic library). Libraries were blocked with 2000 pmol of Singular Blocking oligos and human Cot-1 DNA, then captured with IDT's xGen Exome Hyb Panel v2 overnight (16 hours). After washing, captured libraries were amplified with KAPA HiFi HotStart ReadyMix (10 cycles) and purified. Library quality was assessed on TapeStation and quantified with Qubit 1x HS dsDNA assay.

Libraries were individually clustered at 20 pM and sequenced with a 2x155 run format.

READ ALIGNMENT AND GENERATION OF SEQUENCING QUALITY METRICS

Read alignment and duplicate marking were accomplished via bwa mem (v0.7.15) and GATK4 MarkDuplicates, respectively, implemented using Nvidia Parabricks (v3.7.0-1) pbrun fq2bam command. A distance of 300 units (approximately 1.4um) was used to mark optical duplicates. The reference consisted of GRCh38 build with decoy contigs used as part of the 1000 Genomes Project and downloaded from: (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome). Hybrid-selection metrics were calculated using GATK4 CollectHSMetrics and target/bait region bed files obtained from the manufacturer.

CREATION AND VALIDATION OF A CUSTOM DEEPVARIANT MODEL

Replicate exome sequencing of HG002-HG006 was performed using either 2x100bp or 2x150bp reads to achieve a mean target read depth of ~100x. DeepVariant (v1.3.0) model training, testing, and validation was performed with the following parameters, epochs = 5, batch_size = 128, learn_rate = 5e-4, alt_align = 'rows', min_SNP_allele_fraction = 0.12, min_indel_allele_fraction = 0.03. Model training was performed on chromosomes 1-19 of the replicates using a warmstart from the Illumina WGS model. Model testing was performed on chromosome 21 of the replicates.

The DeepVariant model performance was validated by applying the trained and tested model to identify variants in an HG001 exome library. Performance over target regions was assessed using hap.py (v0.3.12, https://github.com/Illumina/hap.py) and the GIAB truth vcf (v4.2.1) obtained from https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/GRCh38/.

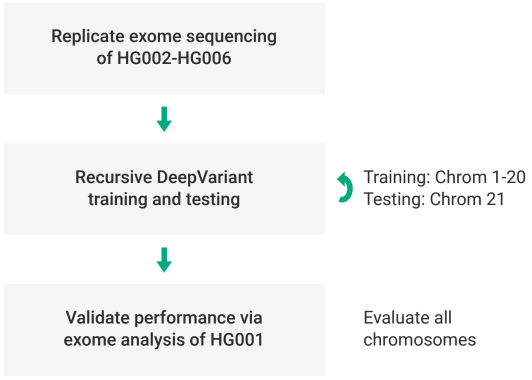
Replicate exome sequencing of HG002-HG006
Recursive DeepVariant training and testing
Training: Chrom 1-20
Testing: Chrom 21
Validate performance via exome analysis of HG001
Evaluate all chromosomes

Figure 1: 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.

Results

Four exome libraries were prepared for GIAB (Genome in a Bottle) samples HG001-HG004 using the IDT KAPA exome kit, followed by 2x150bp sequencing via the F2 flow cell (150M reads). Sequencing metrics are shown in **Figure 2**.

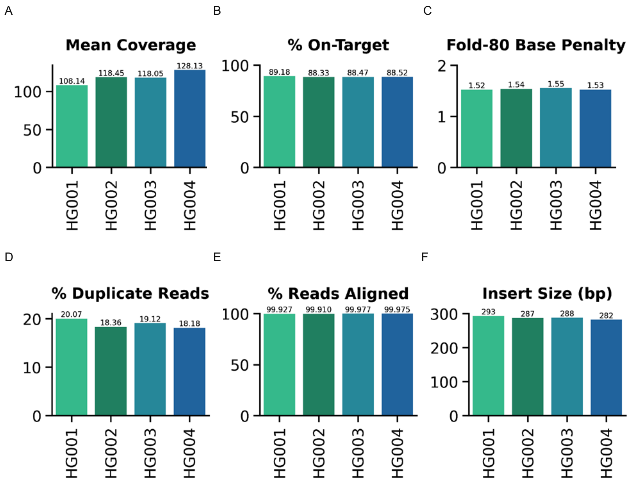

Figure 2: Sequencing quality metrics. Picard tools was used to determine the mean target coverage, percent on-target reads, fold-80 base penalty, percent duplicate reads, percent aligned reads, and mean insert size distribution for each library (A-F, respectively). HG001 data was used to validate performance. All data met system quality specifications: 88 and 80% Q30 for R1 and R2, respectively; accuracy >99.7%.

High coverage unformity was seen across the exome target regions with minimal GC bias as shown in **Figure 3**.

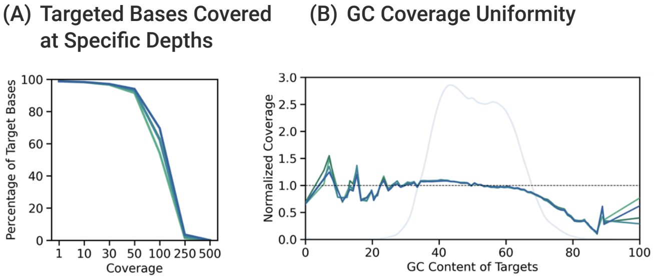

Figure 3: Coverage and GC bias metrics. (A) Read coverage across targeted bases. Coverage values are derived from Picard tools CollectHSMetrics. (B) Coverage uniformity as a function of GC content. Values represent the relative read coverage over panel target regions of a given GC content, normalized to the mean coverage across all target regions. Gray line indicates frequency of targets by GC content.

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). Performance was assessed using hap.py with the NIST GIAB v4.2.1 truth set. Performance metrics derive from hap.py, as seen in **Table 1**.

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: Germline variant detection metrics for HG001.

Conclusion

We have produced a high performing custom DeepVariant model for exome analysis on the G4 Sequencer. 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, but also the suitability of the sequence data for variant detection applications.

*FASTQ files from this study are available by request for additional analysis.

REFERENCES

1. Suwinski, P. *et al*. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Front. Genet.* **10**, 49 (2019).
2. Backman, J. D. *et al*. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
3. Poplin *et al*. BioRxiv (2018) doi: 10.1101/092890
4. Kumaran *et al*. BMC Bioinf (2019) doi: 10.1186/s12859-019-2928-9
5. Koch, L. The power of large-scale exome sequencing. *Nature Reviews Genetics* vol. 22 549 (2021).

SINGULAR
GENOMICS

Get in Touch with the Support

The purchase of a G4 comes with the assistance of a world-class experienced team, consisting of industry veterans, to help you every step of the way. Our customer care team will assist you with order placement and can address any questions you may have. Our field service engineers (FSE) ensure a successful installation and provide instrument support and our field application scientists (FAS) conduct training and validation of your desired application. Our team is committed to support you when you need us.

Whole exome sequencing data is available by request. Please contact customer support at care@singulargenomics.com.

Your experienced team is comprised of:

Customer Care Specialists • • • • **Field Application Scientists** • • • • **Field Service Engineers**

Begin Your Journey with G4

Contact our sales team to learn more about how the G4 can transform your sequencing workflows.

Website: www.singulargenomics.com
Email: care@singulargenomics.com
Call: +1 442-SG-CARES (442-742-2737)
Address: 3010 Science Park Rd, San Diego, CA 92122

Research Use Only. Not for use in diagnostic procedures.

© Singular Genomics Systems, Inc. All rights reserved. G4 and G4 Sequencing Platform are trademarks of Singular Genomics Systems, Inc.