SINGULAR GENOMICS

APPLICATION NOTE

# Whole Exome Sequencing on the G4®

- Rapid sequencing-by-synthesis (SBS) enables cost-efficient delivery of 6–52 exome samples in less than 24 hours.
- The G4 Sequencing Platform fits into existing whole exome sequencing 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 polymorphism (SNPs), 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 on 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 on 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 whole exome sequencing applications. G4 leverages a novel, 4-color Rapid sequencing-by-synthesis (SBS) chemistry to deliver highly accurate reads (single or paired-read format with optional index reads) in less than 24 hours. 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.

# G4 Whole Exome Sequencing Parameters

The G4 Sequencing Platform supports two different flow cell densities, the F2 flow cell and the F3 flow cell. G4 features a 4–flow cell design, enabling users to run 1, 2, 3, or 4 flow cells at a time. Each flow cell has 4 independent lanes, enabling up to 16 independent lanes per run, providing users flexibility in designing sequencing experiments. Sequencing output, run time, quality, and exome throughput by flow cell and run are shown in **Table 1**.

| Flow Cell Type | F2 | F3 |
|---|---|---|
| Read Length (bp) | Up to 2x150 | |
| Run Time (Hours)[a] | 19–24 | |
| Reads / Flow Cell | 200M | 400M |
| Reads / Run | 800M | 1.6B |
| Quality | >85% bases ≥ Q30 | |
| Samples / FC[b] | 6 | 13 |
| Samples / Run[b] | 24 | 52 |

**Table 1** Exome sequencing parameters.
[a]Run time is dependent on read length.
[b]Assumes 34 Mb at 100x coverage.
Throughputs listed are approximations and not guaranteed above kit specifications. Results may vary based on experimental design and sample type. Contact Support for more details.

# Methods

### NGS Library Preparation and Sequencing

Using QuantaBio's sparQ DNA Frag & Library Prep Kit, 150 ng human genomic DNA HG001-HG004 (NIST) was enzymatically fragmented then inactivated. Singular Genomics Universal Adapters were ligated (0.075 uM final adapter concentration), activated, and purified with sparQ PureMag beads. Libraries were amplified with QuantaBio's HiFi PCR Master Mix and Singular Genomics PCR 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 2,000 pmol of the Singular Genomics 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 2x150 bp 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 can be downloaded here. 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 2x100 bp or 2x150 bp 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) and the GIAB truth vcf (v4.2.1).

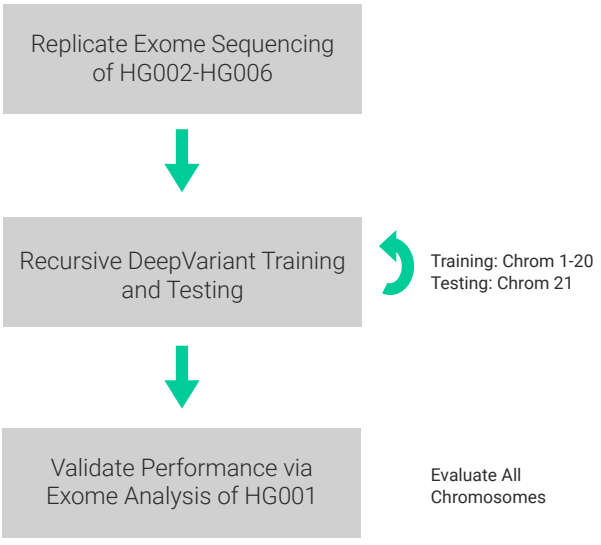


**Figure 1** Workflow for training of a custom DeepVariant model. Data from HG002-6 was used for recursive model training using a warm start from the Illumina whole genome model. Finally, model performance was validated using HG001 exome data.

# Results

Four exome libraries were prepared for GIAB (Genome in a Bottle) samples HG001–HG004 using the IDT KAPA exome kit, followed by 2x150 bp sequencing via the F2 flow cell. Sequencing metrics are shown in **Figure 2**.

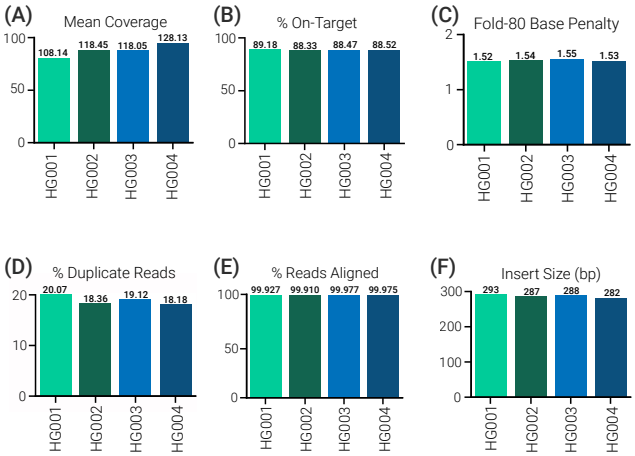


**Figure 2** Sequencing quality metrics. Picard tools was used to determine the mean target coverage, percent on-target reads, fold-80 base penalty, percent duplicate reads, percent aligned reads, and mean insert size distribution for each library (A-F, respectively). HG001 data was used to validate performance.

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

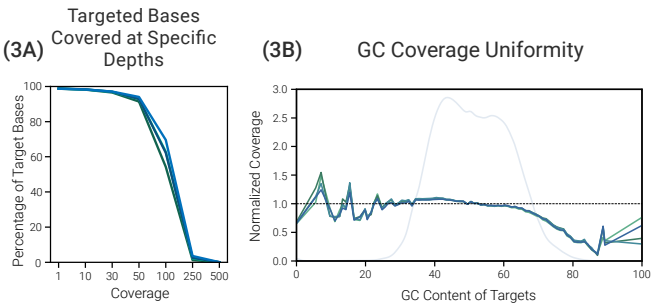


**Figure 3** Coverage and GC bias metrics. **(A)** Read coverage across targeted bases. Coverage values are derived from Picard tools CollectHSMetrics. **(B)** Coverage uniformity as a function of GC content. Values represent the relative read coverage over panel target regions of a given GC content, 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 2x150 bp 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 2**.

| 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 (<50 bp) Precision | 97.45% | 97.76% |
| Indel (<50 bp) 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 2** Germline variant detection metrics for HG001.

# Conclusion

We have produced a high performing custom DeepVariant model for exome analysis on the G4 Platform. 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 and 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 Sequencing Platform with common exome library preparation kits and the suitability of the sequence data for variant detection applications.

SINGULAR GENOMICS

# Get in Touch with the Customer Care Team

The purchase of a G4 comes with the support 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 system installation and provide instrument service and support. Our field application scientists (FAS) conduct user training and provide support for your applications. Our team is committed to support you when you need us.

**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 the capabilities of the G4 Sequencing Platform.

**Website:** www.singulargenomics.com
**Email:** care@singulargenomics.com
**Call:** +1 442-SG-CARES (442-742-2737)
**Address:** 3010 Science Park Rd, San Diego, CA 92121


REFERENCES

1. Suwinski et al. Advancing personalized medicine through the application of whole exome sequencing and Big Data Analytics. Frontiers in Genetics. 2019;10. doi:10.3389/fgene.2019.00049
2. Backman et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021;599(7886):628–34. doi:10.1038/s41586-021-04103-z
3. Poplin et al. Creating a universal SNP and small Indel variant caller with Deep Neural Networks. 2016; doi:10.1101/092890
4. Kumaran et al. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics. 2019;20(1). doi:10.1186/s12859-019-2928-9
5. Koch L. The power of large-scale exome sequencing. Nature Reviews Genetics. 2021;22(9):549–549. doi:10.1038/s41576-021-00397-x


Research Use Only. Not for use in diagnostic procedures.