SINGULAR GENOMICS

TECHNICAL REPORT 2.0

Human Genome Sequencing

Performance Characterization of the G4® Sequencing Platform for Human Whole Genome Sequencing

Introduction

Next-generation sequencing (NGS) has become a foundational tool for both biological research and *in vitro* diagnostics, particularly in oncology, immunology, and detection of genetic disorders. Despite its success, limitations of traditional NGS systems include long analysis times, labor intensive protocols and extensive sample batching requirements for cost-effective use.

To address these limitations, Singular Genomics developed the G4 Sequencing Platform for rapid and flexible sequencing. The G4 delivers highly accurate reads with flexible throughput in under 24 hours, enabled by a novel reversible terminated nucleotide sequencing-by-synthesis (SBS) chemistry and state-of-the-art optics. Here, we apply the high density F3 flow cell to perform whole genome sequencing of the human reference cell line HG002. With approximately 200M reads per F2 flow cell and 400M reads per F3 flow cell, the F3 doubles the throughput of the F2 flow cell without compromising data quality or turnaround time. Using four F3 flow cells in parallel, customers may achieve 30x coverage of four human genomes in under 24 hours with a single G4 sequencing run.

Methods

NGS Library Preparation and Sequencing

1 µg of gDNA from HG002 (National Institute of Standards and Technology, Cat #RM8391) was used for PCR-free library preparation via the SparQ DNA Library Preparation Kit (Cat #95191-096) with Covaris shearing. Cleavable stem-loop adapters comprising a flow cell priming sequence and a sequencing primer sequence were ligated to the gDNA fragments at a final concentration of 0.75 µM. A cleavage enzyme was added into the SparQ kit after ligation. The library was quantified via qPCR and 1 µL of sample was PCR amplified to determine library size using an Agilent TapeStation 40 pM of library was used for onboard amplification of one F3 flow cell followed by 2x150 bp sequencing. The workflow steps between library loading and completion of SBS are fully automated on the G4 Platform. Results presented in this report were obtained on a production version of the G4, which allows for a less than 3-min cycle time, and under a 24 hour turnaround time for 2x150 bp sequencing, including integrated cluster generation.

Four-color fluorescent images of clusters on the patterned flow cell were analyzed by an image processing algorithm to extract signal values in each sequencing cycle, followed by application of a base-calling algorithm that determines the base and assigns a quality score based on the confidence of the base-call. Raw reads were subjected to default quality filtering to eliminate polyclonal clusters and any low-quality clusters.

Read Alignment and Generation of Sequencing Quality and Mapping Metrics

Read alignment and duplicate marking were accomplished via bwa mem (v0.7.15) and GATK4 MarkDuplicates, respectively. Alignment was performed using the UCSC GRCh38 golden path analysis set, and a distance of 500 pixels (approximately 1.6 um) was used to mark optical duplicates. These steps were implemented using either NVIDIA Parabricks (v4.0.1-1) pbrun fq2bam command for quality metrics or using the native tools with a soft-clipping penalty included for variant calling, as this parameter is not implemented in Parabricks. Base and read quality metrics were calculated using GATK4 CollectQualityYieldMetrics. Coverage and alignment percentages were calculated using GATK4 CollectWgsMetrics. Median insert size was calculated using GATK4 CollectInsertSizeMetrics. Normalized GC Coverage was calculated using GATK4 CollectGcBiasMetrics. To calculate the rate of each error mode as a function of cycle number and GC content, a modified reference was created to account for known variants in HG002. This was accomplished by using GATK4 FastaAlternateReferenceMaker with the reference described above and the GIAB HG002 truth vcf (v4.2.1). Reads were aligned to this modified reference, and per GC and per cycle error rates were calculated as described in (6) with exclusion of base calls with a quality score less than 10.

Variant Calling and Processing

A custom DeepVariant (v1.4.0) model was built by warm-starting from the Illumina® WGS model and training on previous G4 sequencing runs of an HG001 library. Training included the optional insert size channel as well as alt-align rows. Germline variants were then called on HG002 using DeepVariant (v1.4.0), implemented via Parabricks (v4.0.1-1) pbrun deepvariant, with a Parabricks-compatible converted version of the custom model and the following parameters: min-mapping-quality 10, min-base-quality 5, min-snp-frac 0.12, min-indel-frac 0.005, alt-align rows, optional-channels insert-size. The standard and Parabricks-compatible versions of the DeepVariant custom model is available upon request on the Singular Genomics website. Performance for high confidence region variants was assessed using hap.py (v0.3.12) with GIAB high confidence regions and truth vcf (v4.2.1).

Results

To assess performance of the F3 flow cell, we prepared a human whole genome sequencing library from 1 µg of Covaris-sheared gDNA from the human reference control HG002 (methods). Sequencing was performed using a single F3 flow cell with 2x150 bp reads format, yielding a total of 413,834,994 read-pairs, for a mean coverage of 33.6x 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.[2]

Read quality and accuracy were high (88.6% and 92.6% of base calls ≥ Q30, **Figure 1A, 1B**; mean single-pass accuracies of 99.87% and 99.92%, Read 1 and Read 2 respectively). Per-base quality estimates closely correlate with empirically determined error rates, allowing for reliable base-call confidence prediction, and suggesting the sequencing run quality may be assessed directly from FASTQ files without the need for post-sequencing base quality score recalibration **(Figure 1C)**.

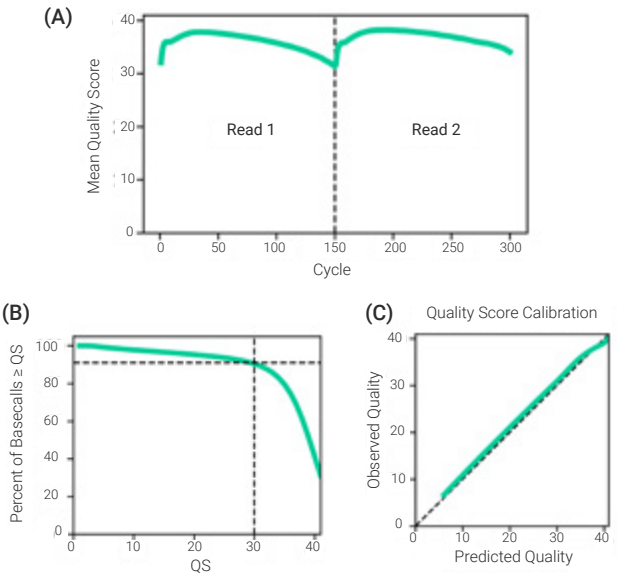

Figure 1 (A) Mean quality score (QS) as a function of cycle for Read 1 and 2. **(B)** Cumulative base quality plot indicating the percentage of base calls (y-axis) having a predicted quality score greater than or equal to value reported on the x-axis. Figure reflects aggregate of Read 1 and Read 2. **(C)** Observed versus predicted base quality, expressed as Phred scores: -10*log10(error rate).

Following read alignment, coverage uniformity was assessed by evaluating the genome-wide distribution of mapped reads as a function of GC content, as measured by Picard CollectGcBiasMetrics **(Figure 2)**. Normalized coverage remains within 0.75 to 1.5 over a wide range of GC content, matching the pattern observed on other leading NGS platforms.[3]

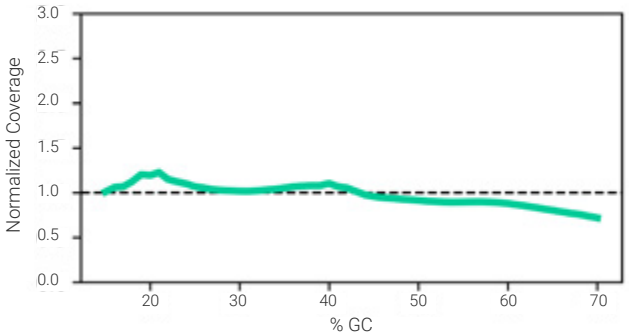

Figure 2 Normalized coverage as a function of GC content. Normalized coverage was calculated as the mean coverage for all 100 bp windows in the genome having a given GC content, divided by the mean genomic coverage. Coverage is shown for a range of GC content representing 99.9% of the human genome.

Further, the coverage distribution closely fits the theoretical distribution of a random process (*i.e.*, Poisson distribution), indicating that coverage is largely unaffected by GC content or other genomic features **(Figure 3A, 3B)**. We observe broad coverage of the human genome, with 96.5% of total genome bases and 99.5% of NIST GIAB v4.2.1 high confidence region bases covered by at least ten reads **(Table 1)**. Notably, we observe an excellent insert length distribution (median 328 bp, **Figure 4**), an important factor driving the low duplication rate observed in this dataset.

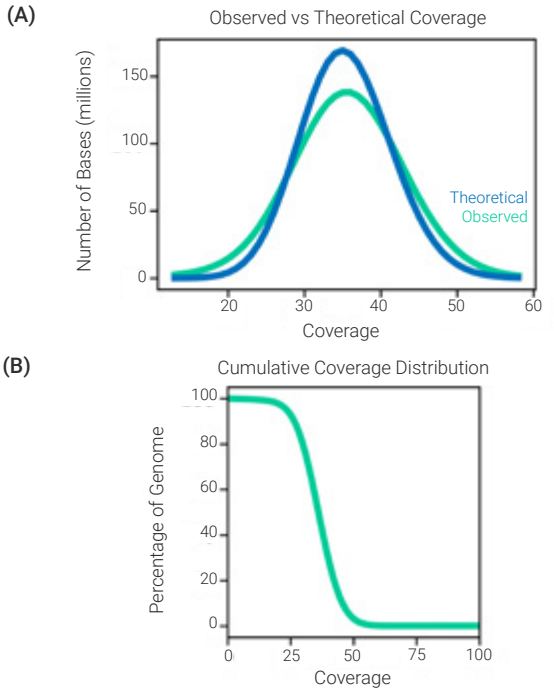

Figure 3 (A) Observed (green) vs theoretical (blue) coverage uniformity in HG002. Values reflect coverage over high confidence regions of the genome. **(B)** Cumulative distribution of per-base coverage over high confidence regions of the genome.

Metric	Value
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 (<50 bp) Precision	98.33
Indel (<50 bp) Recall	97.43
Indel F1-Score	97.88
Total SNPs	3,755,346
Het:Hom Ratio	1.51
Ti:Tv Ratio	2.00

Table 1 Sequence quality and variant detection performance metrics. Variant detection was performed using DeepVariant v1.4 and a custom model. Performance metrics, including the Het:Hom ratio and Ti:Tv ratio, were obtained via hap.py as described in Methods.

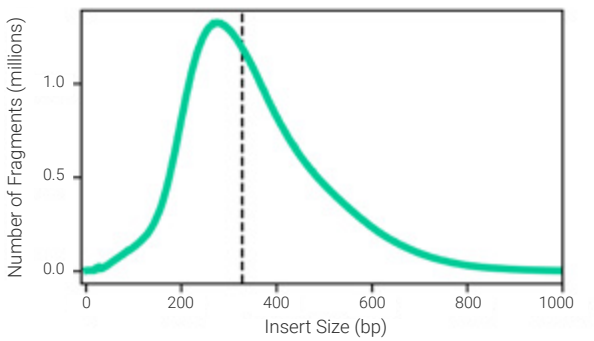

Figure 4. Insert size distribution following read mapping to the genome via bwa mem. Vertical dash indicates median insert size.

In SBS with reversible terminator nucleotides, the dominant error mode is typically substitution errors, while insertion and deletions errors are quite rare.[4] To assess error modes, we examined read accuracy as a function of cycle and GC content (**Figure 5A, 5B**, respectively). We observed a high single-pass accuracy across all sequencing cycles and over a wide range of GC content, with substitution errors dominating over insertion and deletion errors. The similarity of G4 sequencing error modes to that of other SBS platforms enables a high compatibility with existing bioinformatic tools, minimizing the need to develop custom algorithms for data analysis and interpretation, as has now been demonstrated by numerous collaborator studies across varied NGS applications (available for download on the Singular Genomics website).

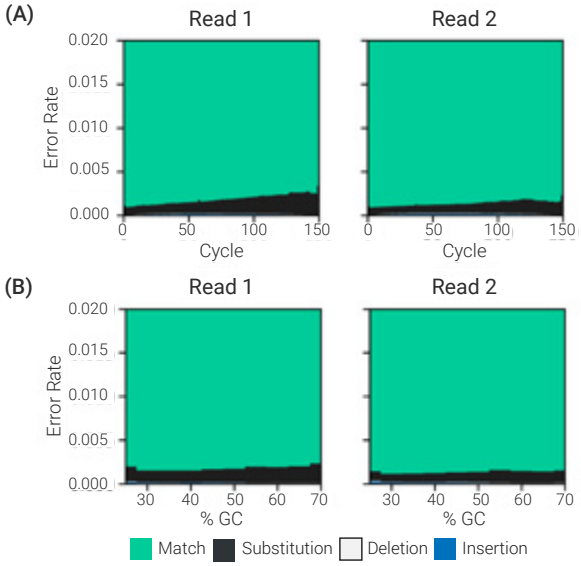

Figure 5 Rate of each error mode for each of the four flow cells as a function of **(A)** cycle and **(B)** GC content.

Finally, to further examine data quality, we performed germline variant analysis of aligned data using a custom-trained DeepVariant[5] model optimized for the G4 Platform.[6] We observed high precision and sensitivity across all variant types within high confidence regions of the genome, similar to typical reported values at an equivalent depth of coverage[7,8] **(Table 1)**.

Conclusion

Sequencing of a highly characterized human reference genome serves as an advanced benchmark for assessment of DNA sequencing performance, including metrics such as single-pass accuracy, throughput, reproducibility, coverage uniformity, and variant calling accuracy. The G4 Sequencing Platform produced sequencing data on par with current state-of-the-art NGS performance, with single-pass accuracy of ~99.9%, and uniform coverage of the high-confidence regions in the reference genome, all while delivering a rapid turnaround time and flexible throughput.

While whole human genome sequencing offers a rigorous test of performance, some of the capabilities of the G4 Platform are particularly well-suited to targeted applications, such as exome sequencing, gene panels for tissue and liquid biopsy, immune repertoire analysis, and single-cell gene expression profiling. Excellent G4 performance in these applications has been demonstrated through numerous internal and collaborative studies with library preparation partners, available for review through the Singular Genomics website. We envision the features enabled by this platform–rapid run time, high read accuracy, scalable sequencing capacity, and independent handling of samples in separate flow cell lanes–in combination with the higher throughput of the F3 flow cell, will have broad applications in biological research and translational medicine.

Sequencing data used within this Technical Report is available by request.

Begin Your Journey with G4

Contact our sales team to learn more about the capabilites 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. Zook *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data. 2016;3(1). doi:10.1038sdata.2016.25
2. Picard [Internet]. [cited 2023 Sept 22]. Available from: http://broadinstitute.github.io/picard
3. SPARQ DNA Frag & Library Prep Kit: Next generation sequencing [Internet]. 2023 [cited 2023 Sept 22]. Available from: http://www.quantabio.com/product/sparq-fraglibrary-prep
4. Schirmer *et al.* Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17(1). doi:10.1186/s12859-016-0976-y
5. Poplin *et al.* A universal SNP and small-indel variant caller using Deep Neural Networks. Nature Biotechnology. 2018;36(10):983–7. doi:10.1038/nbt.4235
6. Gouin et al. Rapid whole genome and whole exome variant detection using a novel fluorescently labeled reversible terminated nucleotide sequencing system and GPU-based accelerated analysis. ASHG 2022; Poster #A-2180. www.singulargenomics.com/assets/ASHG