SINGULAR GENOMICS

TECHNICAL REPORT 1.0

Human Genome Sequencing

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

INTRODUCTION

The genomic tools and technologies developed since the first sequencing of the human genome have greatly improved the understanding of biology, empowered the development of novel therapies, and advanced clinical diagnostics. 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, current limitations of NGS systems include long analysis times, labor intensive protocols, extensive sample batching requirements for cost-effective use, and limited choice in instrumentation. There is a need for new DNA sequencing platforms that combine high accuracy, speed, and flexible throughput to provide timely results and cost-effective operation for research and clinical applications.

Here we present whole human genome sequencing data from the G4™, a new NGS platform from Singular Genomics, designed to enable accurate sequencing and faster results to advance the state of the art in clinical and basic research. We characterize the platform performance by sequencing a reference human genome, cell line NA12878, from the CEPH Utah Reference collection[1]. Accurate sequencing of the whole human genome, containing approximately 3 billion base pairs, represents a major milestone in the development of a new sequencing technology, and provides an opportunity for in-depth assessment of performance.

Today, the majority of DNA sequencing data comes from sequencing by synthesis (SBS), in which a complementary DNA strand is synthesized using fluorescently labeled nucleotides with reversible terminators. As each nucleotide is added, its identity is determined by imaging the fluorescence emission. The fluorescent dye and the reversible terminator are removed, and the cycle is repeated to produce sequencing reads of typically ~ 150 bases. An important component of the approach is the creation of densely packed, but spatially distinct "clusters" of identical copies of the DNA to be sequenced (typically ~ 100 – 10,000 copies per cluster), which allows for much greater signal-to-noise ratio than single-molecule detection and enables rapid, highly parallelized imaging readout using high resolution microscopy.

The G4 takes advantage of the fundamental strengths of the SBS approach but is engineered from the ground up using novel chemistry, sequencing enzymes, polymer scaffolds, micro-fabrication techniques and molecular biology workflows, in addition to engineering innovations in high-speed imaging and rapid fluid exchanges during sequencing cycles. Besides focusing on key performance metrics – accuracy, speed, throughput – the G4 is designed to be compatible with existing lab ecosystems, including upstream NGS library preparation solutions and downstream bioinformatic pipelines. Sequencing is performed independently on 1, 2, 3 or 4 flow cells in parallel, with each containing 4 individually addressable lanes capable of running separate samples. To maximize flexibility, flow cells with two different output levels will be available on the G4: the F2 flow cell will produce approximately 150M-165M read pairs, and the F3 flow cell will produce approximately 300M-330M read pairs (available in late 2022). Utilizing all four flow cells, a single run with four F3 flow cells will produce approximately 1,200M-1,300M read pairs.

RESULTS

To evaluate system performance and reproducibility, we carried out two independent whole genome sequencing runs in paired-read 2x150 cycle format, each with two F2 flow cells, using the human reference control NA12878. Combined, throughput yielded 692M read-pairs (mean of 173M read-pairs per flow cell; 46X mean coverage when discounting duplicates (15.4%), overlapping portion of reads (15.6%), ambiguously mapped reads (1.0%) and low-quality base calls (0.6%) as reported by Picard[2] CollectWgsMetrics). The combined runs averaged 89% and 91% of bases at qualities greater than or equal to a predicted Phred Score of Q30 for Read 1 and Read 2, respectively (Table 1, Figure 1B).

The accuracy was consistently high across flow cells for each 150-cycle read, with mean single-pass accuracies of 99.82% and 99.84%, Read 1 and Read 2 respectively (Figure 1A), with 85% of Read 1 and 87% of Read 2 reads having 0 errors out to 150 cycles. Basecalls with predicted quality < Q10, representing < 1% of the data, were not considered in the accuracy calculations. 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 1B and 1C).

METRIC	Flow Cell 1	Flow Cell 2	Flow Cell 3	Flow Cell 4
Configuration	2x150	2x150	2x150	2x150
Paired-Reads (M)	168	169	169	186
Output (Gb)	51	51	51	56
% Bases ≥ Q30 R1	90	90	89	88
% Bases ≥ Q30 R2	91	91	90	90

Table 1: Sequencing run metrics across 4 flow cells.

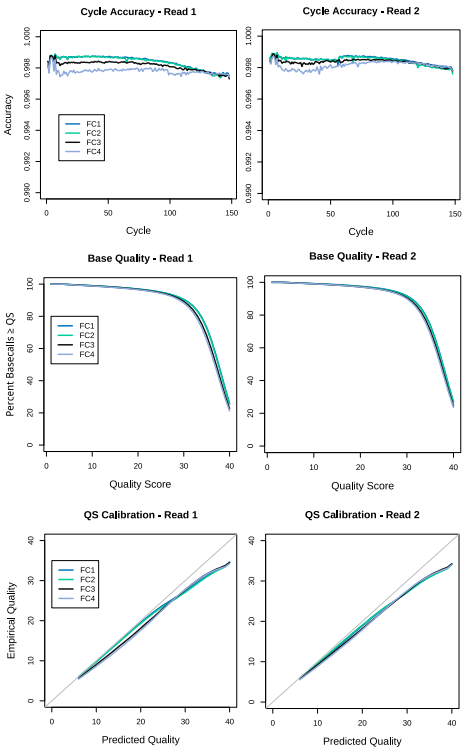

Figure 1:
A). The accuracy of base calls for each cycle are shown for each of the 4 flow cells run in this study (FC1-FC4) for both Read 1 (left plot) and Read 2 (right plot).

B). Cumulative base quality plots show what percentage of base calls (y-axis) having a predicted quality score greater than or equal to value reported on the x-axis.

C). Empirical vs predicted base quality. The predicted and empirical error rates are expressed as Phred scores: $-10*\log_{10}$(error rate).

Following read alignment, coverage uniformity was assessed by evaluating the genome-wide distribution of mapped reads. The observed distribution is consistent with measurement of GC coverage evenness via Picard CollectGcBiasMetrics (Figure 2). Coverage 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 and B). We observed uniform coverage of the human genome, with 93.87% of total genome bases and 99.68% of NIST GIAB v4.2.1 high confidence region bases covered by at least one read (Circos[3] plot Figure 3B).

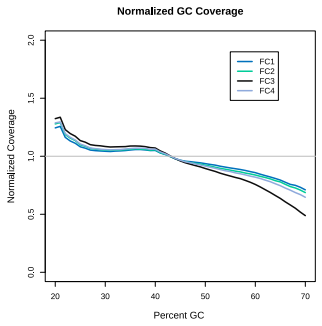

Figure 2: Normalized coverage was calculated as the mean coverage for all 100bp windows in the genome that had the same GC content divided by the mean genomic coverage. A Coverage is shown for a range of GC content that represents 99.9% of the human genome.

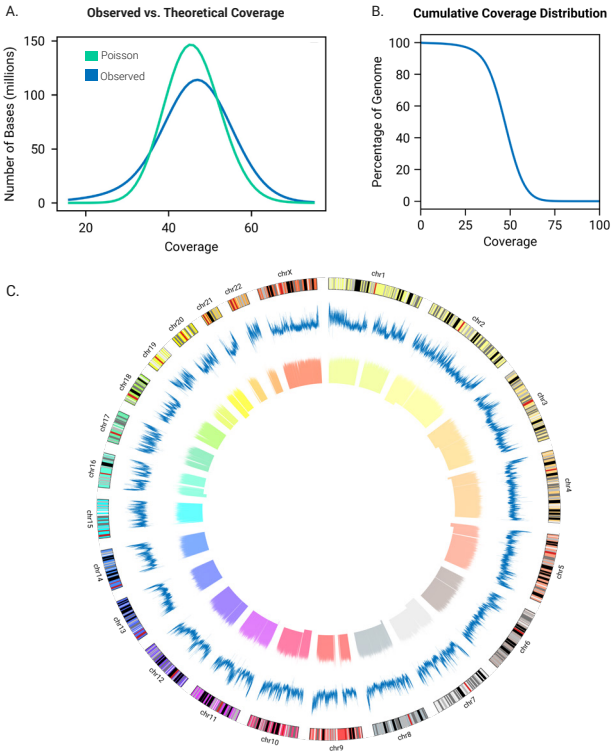

Figure 3: A). Observed vs theoretical coverage uniformity in cell line NA12878 over the combined set of flow cells. **B).** Cumulative distribution of per-base coverage over high confidence regions of the genome over the combined set of flow cells. **C).** Genomic coverage (inner ring, colored by chromosome bar-plot, 1 - 5,889 reads per window) and GC content (middle ring, line-plot, 20% - 70%) in 10kb windows along the genome with chromosomes delineated in the outer ring.

In SBS with reversible terminators, 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 per flow cell as a function of cycle and GC content (Figure 4A and B, 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, as expected.

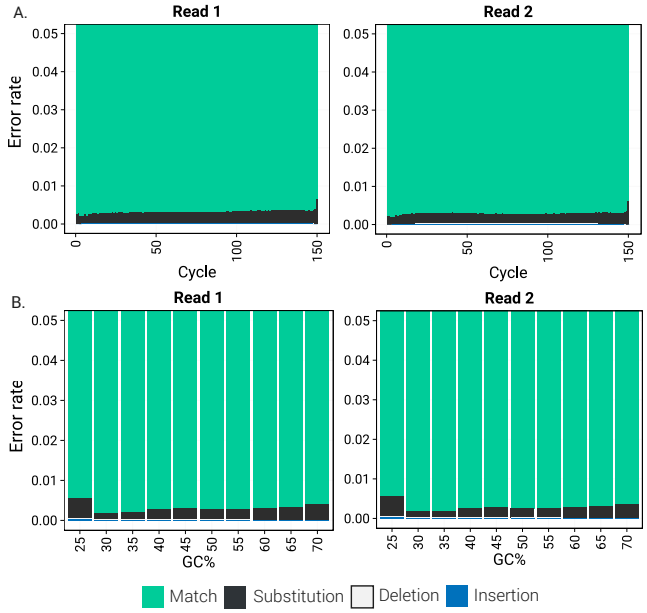

Figure 4: Rate of each error mode for each of the four flow cells as a function of (A) cycle and (B) GC content.

The similarity of G4 sequencing error modes to that of other SBS platforms implies a high compatibility with existing bioinformatic tools, minimizing the need to develop custom algorithms for data analysis and interpretation. To test this possibility, we performed germline variant analysis of aligned data using an implementation of DeepVariant[5] that had been optimized for another common SBS-based platform (Methods). We observed high precision and sensitivity across all variant types within high confidence regions of the genome, similar to typical reported values at equivalent depth of coverage[6,7] (Table 2), even without training the variant caller on data derived from the G4 sequencing platform. Training of DeepVariant and other variant calling algorithms on G4-specific data may result in even greater accuracy.

METHODS

NGS Library Preparation and Sequencing

100 ng of gDNA from NA12878 (National Institute of Standards and Technology - NIST cat# RM-8398) was used for PCR-free library preparation via the SparQ DNA Library Preparation Kit (Cat. 95191-096) with enzymatic fragmentation for 8 minutes. Cleavable stem-loop adapters comprising a flow cell priming sequence and a sequencing primer sequence were ligated to the gDNA fragments. A USER treatment step was added into the SparQ kit after ligation. 20 pM of library was used for onboard amplification of four F2 flow cells (600M combined read capacity) followed by 2x150bp sequencing. The workflow steps between library loading and completion of SBS are fully automated on the G4 instrument. Results presented in this report were obtained on a pre-production version of the G4, operating with a 4-minute SBS cycle time, which is already shorter than most leading NGS systems/kits. The production version of the G4 uses an even faster imaging system which allows for a <3-min cycle time, and a total run time of 16-19 hours for 2x150 bp sequencing, including integrated cluster generation.

METRIC	20x Target Coverage	30x Target Coverage	40x Target Coverage
%PF Reads Aligned	99.99	99.16	98.37
Duplication Rate (%)	17.7	16.3	15.5
Median Insert Size (bp)	251	249	248
Mean Coverage (X)	22.2	33.6	45.9
%Bases >=10x Coverage	98.49	99.44	99.69
SNP Precision	99.30	99.62	99.71
SNP Sensitivity	98.94	99.14	99.17
SNP F1-Score	99.12	99.38	99.44
Indel (<50bp) Precision	95.02	96.49	97.13
Indel (<50bp) Sensitivity	93.51	95.49	96.4
Indel F1-Score	94.26	95.99	96.77
Total SNPs	3738914	3741923	3744535
Het:Hom Ratio	1.49	1.48	1.46
Ti:Tv Ratio	2	1.99	1.99

Table 2: Sequence quality and variant detection at 20x, 30x and 40x target coverage. Variant detection was performed using DeepVariant v1.0. Performance metrics, including the Het:Hom ratio and Ti:Tv ratio, were obtained via hap.py as described in Methods.

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 applying 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, implemented using Nvidia Parabricks (v3.6.1-1) pbrun fq2bam command (https://docs.nvidia.com/clara/parabricks/v3.6/text/germline_pipeline.html). 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). Percent bases > Q30 was calculated using GATK4 CollectQualityYieldMetrics

on Read 1 and Read 2 separately. Coverage and alignment percentages were calculated using GATK4 CollectWgsMetrics. Median insert size was calculated using GATK4 CollectInsertSizeMetrics and normalized GC Coverage was calculated using 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 NA12878. This was accomplished by using GATK4 FastaAlternateReferenceMaker with the reference described above and the GIAB NA12878 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). Stacked bar plots were generated in R using ggplot2 as described in (6). Cycle accuracy and Single Pass accuracy were calculated using fgbio ErrorRateByReadPositon for all base-calls with predicted quality ≥ Q10. Cycles 149 and 150 were excluded from accuracy calculation and plot because BWA cannot confidently distinguish the adapter sequence from sequencing errors in the last 2 cycles of the read. Proportion of perfect reads were calculated only for those reads that mapped with an insert size ≥ 150bp to ensure that the same number of base-calls were made for each cycle.

Variant Calling and Processing

Germline variants were called in NA12878 using DeepVariant (v1.0, --model_type WGS) implemented via Parabricks (v3.6.1-1) pbrun deepvariant, with min-mapping-quality 10 and min-base-quality 10. Performance for high confidence region variants was assessed using hap.py (v0.3.12, https://github.com/Illumina/hap.py) with GIAB high confidence regions (v4.2.1, ref) and the GIAB truth vcf (v4.2.1, ref) obtained from https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/GRCh38/.

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 produced sequencing data on par with current state-of-the-art NGS performance, with single-pass accuracy of ~99.8%, and uniform coverage of the high-confidence regions in the reference genome, evaluated at approximately 20X, 30X and 40X coverage.

While whole human genome sequencing offers a rigorous test of performance, some of the capabilities of the G4 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. 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 – will have broad applications in biological research and translational medicine, particularly in oncology and immunology.

REFERENCES

1. Zook JM, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data 3, 160025 (2016).
2. http://broadinstitute.github.io/picard
3. https://github.com/ponnhide/pyCircos
4. Schirmer et al. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinform. 17, 125 (2016).
5. Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 36, 983–987 (2018)
6. Foox et al. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. Nat Biotechnol. 39, 1129-1140 (2020).
7. Telenti et al. Deep sequencing of ten thousand human genomes. PNAS 42,11901-11906 (2016).

SINGULAR GENOMICS

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

CONTACT

Website: www.singulargenomics.com
Email: care@singulargenomics.com
Call: +1-442-SG-CARES (1-442-742-2737)
Address: 10931 N Torrey Pines Rd, La Jolla, CA 92037