Whole Genome Sequencing on the G4™ Platform with the F3 Flow Cell

Kenneth Gouin III, Mark Kalaj, Liz LaMarca, Sabrina Shore, Timothy Looney, Daan Witters, Eli N. Glezer
[1]Singular Genomics, 3010 Science Park, Rd, San Diego, California

Introduction

Next-generation sequencing (NGS) has achieved widespread adoption as a tool for biological research and in-vitro diagnostics. Despite this success, traditional NGS systems are limited by long analysis times, labor intensive protocols, and the need for extensive sample batching to achieve cost-effective use. To address these limitations, Singular Genomics developed the G4 Platform for rapid and flexible sequencing. Here we apply the novel, higher density F3 flow cell to perform 30x whole genome sequencing of the human reference cell line HG002, in a single flow cell.

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 **(Figure 1).** Users may analyze up to four flow cells of two types (F2: 200M reads[1], F3: 400M reads[1]) in a single run. To facilitate multiplexing, each flow cell comprises four fluidically independent lanes. To assess performance of the F3 flow cell we prepared a PCR-free human whole genome sequencing library from 1µg of Covaris-sheared gDNA from the human reference control HG002.

Figure 1 G4 Platform and Performance Specifications

Power
3.2B Reads per Run
up to 480 Gb per Run

Speed
< 24 Hour Run Times

Flexibility
1-4 Flow Cells per Run
4-16 Lanes per Run

Accuracy
80 - 90% bases
≥ Q30

Results

High Quality 2x150bp Paired Reads

Sequencing via a single F3 flow cell with 2x150bp reads format yielded a total of 413,834,994 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 **(Figure 4).** Insert lengths were varied, with a median of 328bp **(Figure 5)**. Base quality scores were well calibrated **(Figure 6)** and there was minimal GC related coverage bias **(Figure 7).** Consistent with this, the coverage distribution over high confidence regions matches expectation from a Poisson distribution. Error modes were dominated by substitution errors, with insertion and deletion errors comparably rare **(Figure 8).** Accuracy was consistently high over varying GC content ranges.

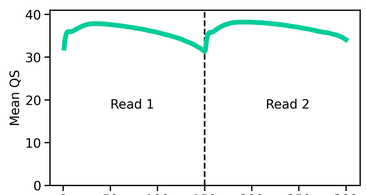

Figure 4 Read quality per cycle

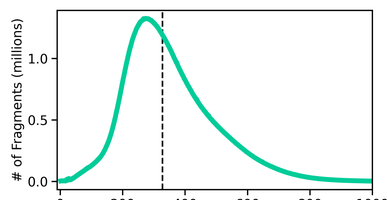

Figure 5 Insert lengths

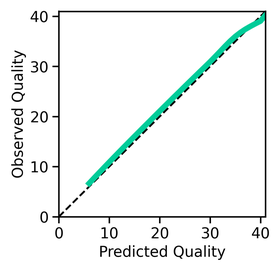

Figure 6 Quality score calibration

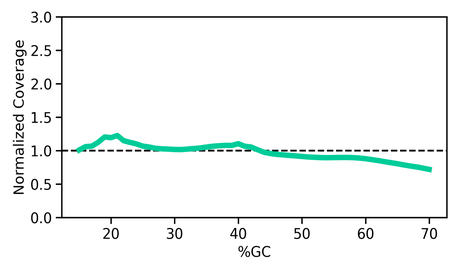

Figure 7 Normalized GC Coverage

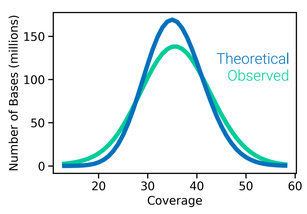

Figure 8 Genome Coverage

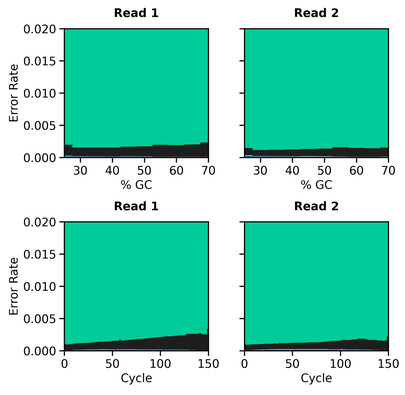

Figure 9 Error modes per cycle and GC content

Results

Germline Variant Detection

To assess germline variant detection performance, we downsampled data to 30x coverage, then identified variants over high-confidence regions of HG002 using a custom-trained DeepVariant[2] v1.4 model deployed on the Nvidia Parabricks platform* . Performance was assessed via hap.py. We observe high precision and recall over high-confidence regions of the genome, similar to typical reported values at an equivalent depth of coverage[3].

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

Table 1 Germline variant detection performance from 30x coverage

Conclusion

The G4 with F3 flow cell produces sequencing data equivalent to the industry standard 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 of up to 1.6 billion paired reads per run. 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.

References

1. http://broadinstitute.github.io/picard
2. Poplin et al. Nat Biotechnol. 36, 983–987 (2018)
3. Telenti et al. PNAS. 42,11901-11906 (2016)

*For detailed methods, raw data, and access to custom trained DeepVariant models, visit www.singulargenomics.com
†Throughputs listed are approximations and not guaranteed above kit specifications. Results may vary based on experimental design and sample type.