Bacterial whole genome assembly and microbial community analysis with the Singular Genomics G4™ platform

Christopher G Daum[1], Jie Wang[1], Laura Sandor[1], Len Pennacchio[1], Sabrina Shore[2], Bubba Brooks[2], Ryan Shultzaberger[2], Timothy Looney[2], Martin M Fabani[2], Eli N Glezer[2]

(1) DOE Joint Genome Institute, Berkeley, California
(2) Singular Genomics Systems, Inc., 3010 Science Park Rd, San Diego, CA 92121

Background

Next generation sequencing (NGS) has become a central component of modern microbiome basic and translational research by enabling high resolution assessment of species diversity and de novo genome assembly at a far lower cost than traditional methods. Here we present our initial evaluation of the novel Singular Genomics G4™ Platform for rapid sequencing by synthesis (SBS). We demonstrate the suitability of the platform for bacterial genome de novo assembly and microbial community analysis.

Methods – G4 Sequencing Platform

The G4 is a benchtop sequencer designed to deliver rapid SBS with throughput flexibility to reduce batching related delays. The G4 supports single or paired end insert reads of up to 150bp, with paired index reads for sample multiplexing.

Figure 1. Microbiome workflow and G4 system. Library preparation involves incorporation of G4 flow cell adapter sequences and optional sample indices.

Methods – Library preparation, sequencing and analysis

Genomic PCR-free libraries were prepared from three bacterial species of varying GC content (E. coli, P. heparinus, M. ruber) and the Zymo mock community control (cat #D63505) using the QuantaBio SparQ library preparation kit, then sequenced on the G4 using an F2 flowcell to produce 178M 2x150 read pairs meeting the G4 specifications (79 and 81% Q30, R1 and R2; accuracy >99.7). Reads were aligned using fq2bam (NVIDIA Clara Parabricks), cycle accuracy was calculated using BBTools Reformat, and quality score calibration and genome coverage was assessed using Picard tools. Bacterial genome assembly was performed with SPAdes v3.14.1 using "--isolate" mode, default parameters and 1M randomly selected read pairs.

Results – Sequencing and genome assembly metrics

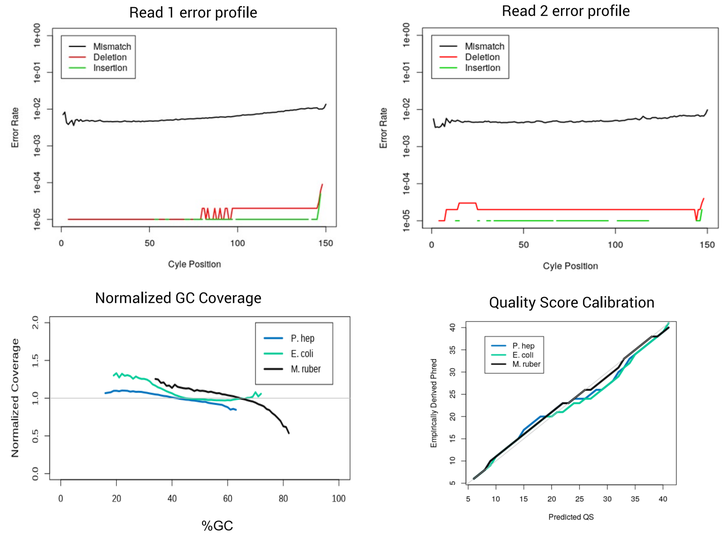

Figure 2. Error profiles for Read 1 and Read 2, quality score calibration, and normalized GC coverage for bacterial genome libraries. GC Coverage plots are restricted to those GC contents observed > 0.01% of all 100bp windows in each respective genome. Error profile derives from E coli library but is representative of the profile observed in each of the three species examined.

Metric	*P. heparinus*	*E. coli*	*M. ruber*
Size (Mb)	5.17	4.64	3.10
GC Content (%)	42.1	50.4	63.40
Avg. Depth	57X	63X	96X
Contigs	45	80	31
Contigs >1000bp	45	80	31
Largest Contig	464714	226908	359730
Total Length	5124777	4650630	3073141
Total Length>1000bp	5124777	4650630	3073141
Genome fraction (%)	99.12%	98.12%	99.13%
N50	240643	107723	181772
Misassemblies	0	1	0
Misassembly length	0	67270	0
Indels per 100kb	0.25	0.15	0.39

Table 1. Assembly metrics for bacterial genome assemblies. Metrics derived from analysis of >1000bp contigs using Quast v5.02

Results – Microbial mock community analysis

Species	Avg. GC (%)	Gram Stain	gDNA Abun. (%)
Pseudomonas aeruginosa	66.2	-	12
Escherichia coli	56.8	-	12
Salmonella enterica	52.2	-	12
Lactobacillus fermentum	52.8	+	12
Enterococcus faecalis	37.5	+	12
Staphylococcus aureus	32.7	+	12
Listeria monocytogenes	38.0	+	12
Bacillus subtilis	43.8	+	12
Saccharomyces cerevisiae	38.4	Yeast	2
Cryptococcus neoformans	48.2	Yeast	2

Table 2. Contents of Zymo mock community standard used in the analysis. The standard contains genomes having a range of GC content. This standard is commonly used in microbiome sequencing to assess library preparation and sequencing quality.

Figure 3. Relative abundance of species within the Zymo mock community as determined by analysis of G4 data. Species abundance was determined by applying Kraken v2.1.2 and Bracken v2.6.2.

Conclusion

In this study we successfully applied the G4 pre-production sequencer to assemble the genomes of three bacteria species spanning a range of GC content, achieving >98% genome coverage with assembled contigs. We then applied the G4 to accurately quantify the abundance of diverse bacterial and fungal species from a mock community sample using off the shelf bioinformatic tools.

The error profile of the G4 is similar to that of other short read platforms, both with respect to the absolute and relative frequencies of substitution and indel errors. Indeed, we find that G4 data may be used with common microbial bioinformatics tools built for other short read platforms without need for modification.

Owing to the relatively modest number of reads required for microbiome analysis and the sensitivity to cost, microbiome sequencing centers must employ large scale pooling of barcoded libraries when sequencing with traditional high throughput platforms, invariably leading to batching related delays. In this respect the flexible throughput of the G4 platform may help reduce turnaround times in basic and translational microbiome research.