Unbiased characterization of metagenome composition and function using HiFi sequencing on the PacBio Sequel II System

Meredith Ashby[1], Greg Young[1], Shreyasee Chakraborty[1], Janet Ziegle[1], Joan Wong[1], Primo Baybayan[1], Richard Hall[1]
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Abstract

The introduction of the Sequel II System has enabled a new, higher throughput, assembly-optional data type, HiFi reads, that addresses common challenges in metagenomics research. HiFi reads combine >99% accuracy with long read lengths, allowing researchers to overcome ambiguous taxonomic assignment in 16S sequencing and eliminating the requirement for data-inefficient assembly for most shotgun sequencing applications, including gene discovery and metabolic pathway reconstruction.

HiFi Sequencing

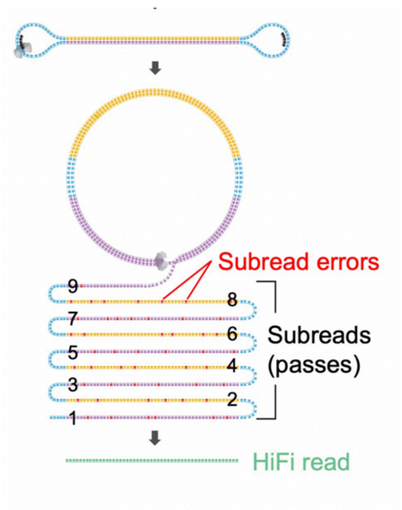
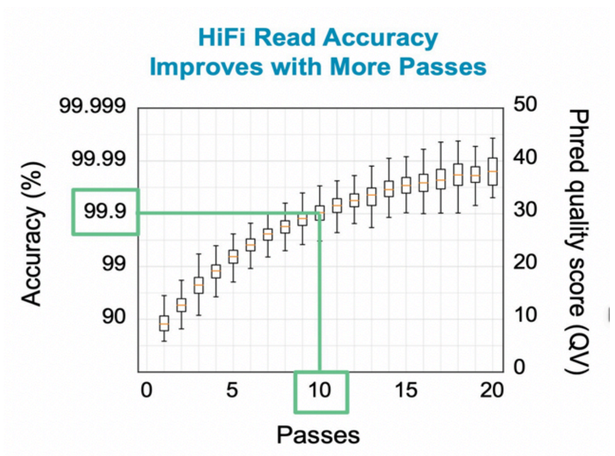


Fig 1. Sequencing advances on the Sequel II System have allowed the *ccs* method to be applied to far longer insert libraries than before. With average read lengths up to 100 kb, both 16S amplicons and 10 kb shotgun metagenomics libraries can be sequenced at very high single-molecule accuracy.



Methods and Data

Table 1. For shotgun profiling, 10 kb SMRTbell libraries were made from Megaruptor-sheared samples, sequenced on the Sequel II System (30 hr), and analyzed with PacBio's *ccs* algorithm.

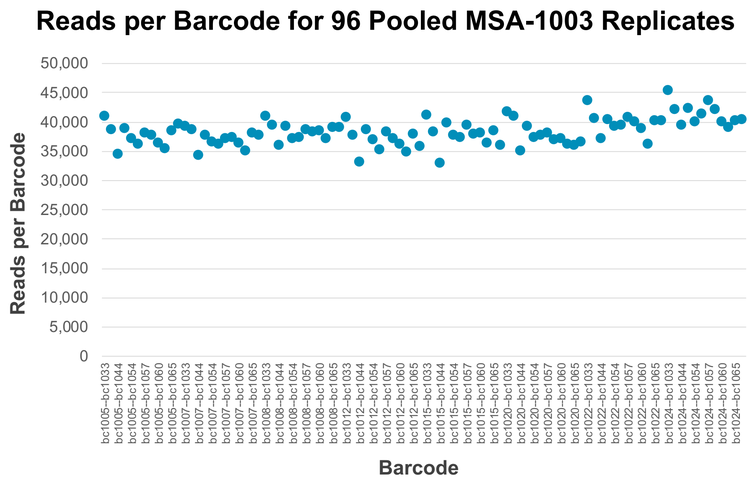
Shotgun	No. >Q20 reads	Avg RL (bp)	>Q20 QV
MSA-1003	2,358,257	8,262	Q35
Human fecal 1	2,485,902	8,806	Q39
Human fecal 2	2,634,276	9,247	Q37
Human fecal 3	2,371,437	8,570	Q39
Human fecal 4	2,133,478	10,104	Q36
Human fecal 5	2,037,230	9,746	Q37
Human fecal 6	2,230,353	8,870	Q39
Human fecal 7	2,796,697	8,120	Q40
Human fecal 8	1,977,870	8,612	Q40
Human fecal 9	2,529,830	8,660	Q39

- In genome sequencing, assembly is typically required to improve either the contiguity or accuracy of raw data.
- Since the abundance of organisms in metagenomes can vary by orders of magnitude, often a significant proportion of the data is from species without enough coverage for this key step.
- Depending on sample type, 20-80% of raw data is typically not incorporated into metagenome assemblies and is lost to further analysis.
- The length and accuracy of HiFi reads match or outperform those of many metagenome assemblies, enabling cost-effective recovery of intact genes and operons even from species with low representation in the data. Every HiFi read is useful for analysis.

Table 2. Full-length 16S data was collected for ATCC® 20-strain even (MSA-1002™) and staggered (MSA-1003™) mock communities on single SMRT Cells 8M at 96-plex. Amplification was performed with a barcoded universal primer / 2-step PCR or a barcoded 16S primer / 1-step PCR approach, respectively.

16S	Total >Q20 Reads	>Q20 Read Quality	Avg Reads / BC
MSA-1003	3,851,493	Q40	38,349

16S Sequence Analysis



Figs 2 and 3. With the single-step PCR protocol, the yield of >99% accurate reads is highly consistent across all barcodes and matches the expected composition of the control sample.

HiFi Metagenomic Analysis

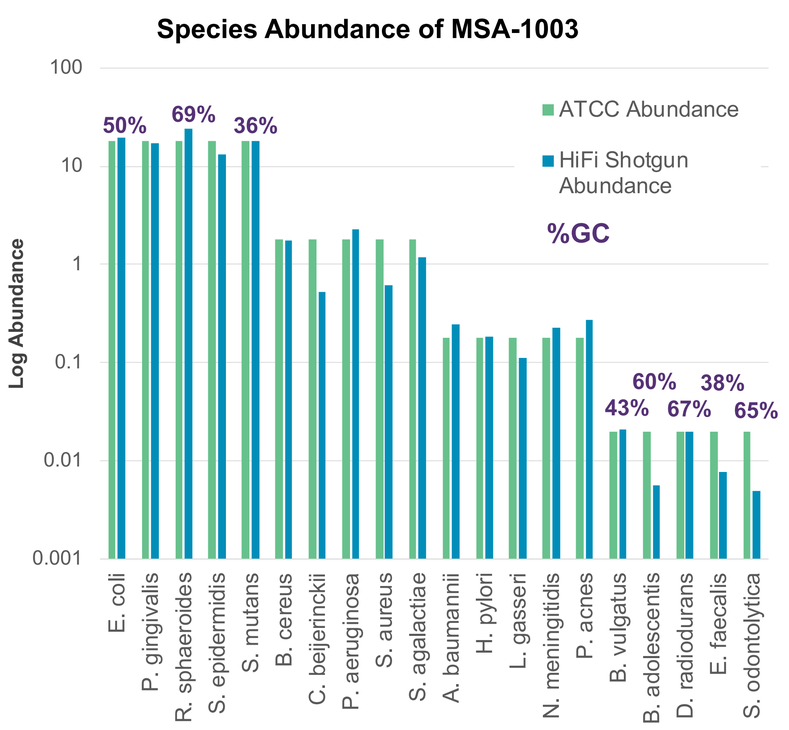


Figure 4. Shotgun sequencing of MSA-1003. Species down to 0.018% abundance were detected successfully in accordance with their expected abundance. The high correspondence between the expected community composition and shotgun profiling data reflects the low context bias of SMRT Sequencing. Download (SRX6095783) to explore the HiFi data set yourself.

Assembly-Free Gene Finding

Table 3. FragGeneScan was used to identify genes in the human fecal samples using unassembled HiFi reads as input. The length of HiFi data is a key advantage for finding intact genes and operons as part of a functional profiling study.

Sample	# Predicted Genes	Mean RL (bp)	Mean Genes / Read	Unique Genes (99% ID)
HF 1	19,639,322	1,005	7.9	1,012,982
HF 2	22,064,417	1,001	8.4	1,141,123
HF 3	18,059,181	1,024	7.6	1,154,341
HF 4	19,844,033	978	9.3	1,250,711
HF 5	18,396,237	970	9.0	1,087,015
HF 6	17,970,195	999	8.1	1,221,052
HF 7	20,773,647	991	7.4	1,119,855
HF 8	15,734,038	977	8.0	862,405
HF 9	19,528,176	1,009	7.7	546,581

Assembly and Binning

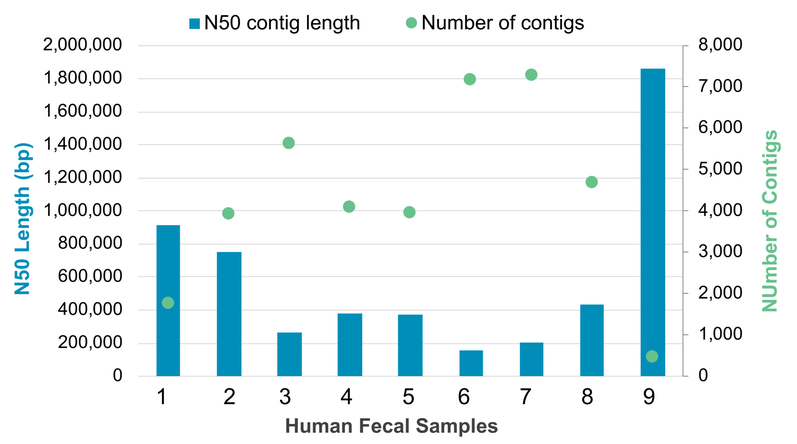


Fig 5. If assembly is desired, for example to create reference genomes from culture-resistant microbes, excellent results can be achieved with Canu and 15- to 20-fold coverage of the species of interest.

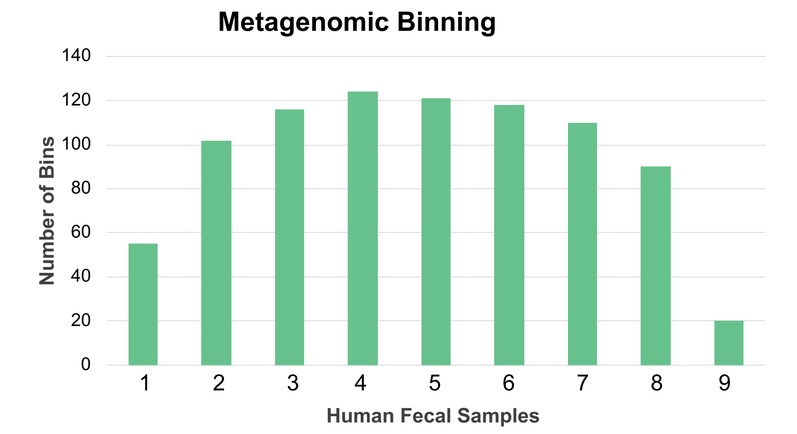


Fig 6. Binning analyses performed with PATRIC / RBS on the combined Canu contigs and unassembled reads for each sample. RBS attempts to reconstruct complete genomes by seeding bins with one contig that encodes a unique 'seed role' protein, associating a reference genome with each bin, then using protein kmers to populate the bins.

Conclusions

- One SMRT Cell 8M yields >30,000 HiFi reads / sample at 96-plex, economically providing species-level community composition information.
- HiFi 16S and shotgun data correspond closely with expected mock community compositions, reflecting the low GC bias of SMRT Sequencing technology.
- HiFi reads yield an average of 7-9 intact, highly accurate genes per read, eliminating the requirement for data-inefficient assembly for metagenome functional profiling studies and providing information even from species with low coverage.
- The high accuracy of HiFi data means it can be analyzed with existing bioinformatic pipelines.