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

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

Abstract

Recent work comparing metagenomic sequencing methods indicates that a comprehensive picture of the taxonomic and functional diversity of complex communities will be difficult to achieve with short-read technology alone. While the lower cost of short reads has enabled greater sequencing depth, the greater contiguity of long-read assemblies and lack of GC bias in SMRT Sequencing has enabled better gene finding. However, since long-read assembly requires high coverage for error correction, the benefits of unbiased coverage have in the past been lost for low abundance species.

SMRT Sequencing performance improvements and the introduction of the Sequel II System has enabled a new, high throughput data type uniquely suited to metagenome characterization: HiFi reads. HiFi reads combine high accuracy with read lengths up to 15 kb, eliminating the need for assembly for most microbiome applications, including functional profiling, gene discovery, and metabolic pathway reconstruction. Here we present the application of the HiFi data type to enable a new method of analyzing metagenomes that does not require assembly.

HiFi Reads on the Sequel II System

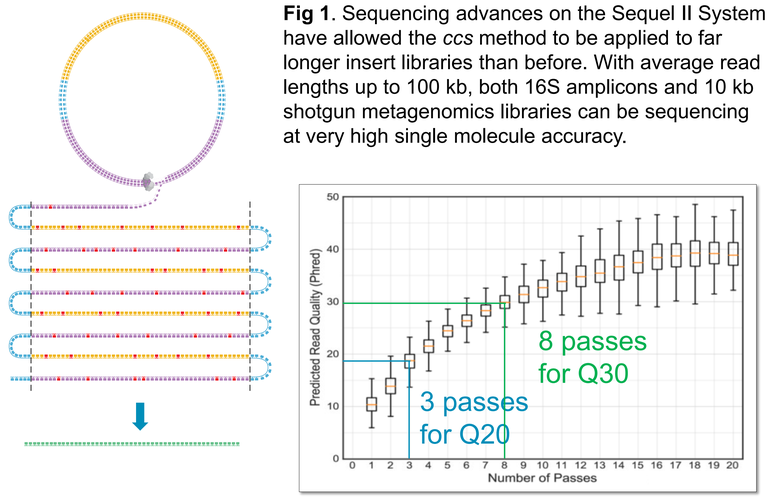

Fig 1. Sequencing advances on the Sequel II System have allowed the *ccs* method to be applied to far longer insert libraries than before. With average read lengths up to 100 kb, both 16S amplicons and 10 kb shotgun metagenomics libraries can be sequencing at very high single molecule accuracy.

Methods and Sequencing Performance

Table 1. Full-length 16S sequencing and shotgun profiling data were collected for mock communities (ATCC® 20 strain staggered (MSA-1002™) and even (MAS-1003™) mixes) For 16S sequencing, V1-V9 amplicons were sequenced on a single SMRT Cell 8M at 48 or 96-plex using either a barcoded universal primer / 2-step PCR approach (MSA-1002) or a barcoded 16S primer / 1-step PCR approach (MSA-1003)

16S	>Q20 BC reads	>Q20 Read Quality	BC Samples / SMRT Cell	Avg reads / BC
MSA-1002	2,426,218	Q42	96	35,000
MSA-1003 (1a)	1,743,260	Q34	48	36,317
MSA-1003 (2a)	1,738,543	Q34	96	18,109

Table 2. For shotgun profiling, SMRTbell libraries with 10 kb inserts were sheared with Megaruptor, sequenced with 30-hour movies on the Sequel II System, and analyzed with PacBio's *ccs* algorithm to obtain HiFi reads with >QV20 accuracy.

Shotgun	>Q20 reads	Avg read length	>Q20 QV
MSA-1003 (1b)	1,976,744	8,454	Q35
MSA-1003 (2b)	2,358,257	8,262	Q35
Human fecal 1	2,802,471	6,289	Q39
Human fecal 2	1,646,208	7,907	Q34
Human fecal 3	1,593,641	8,842	Q34

16S Sequencing

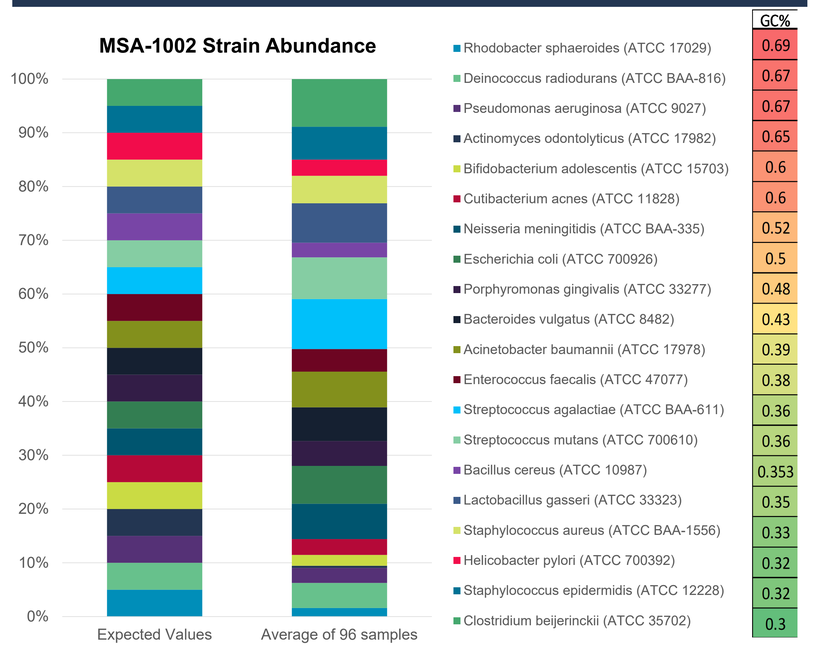

Figure 2. There is high correspondence between the expected and measured abundance of the even-composition mock community, reflecting the low context bias of the sequencing technology. Strains were assigned with BLAST.

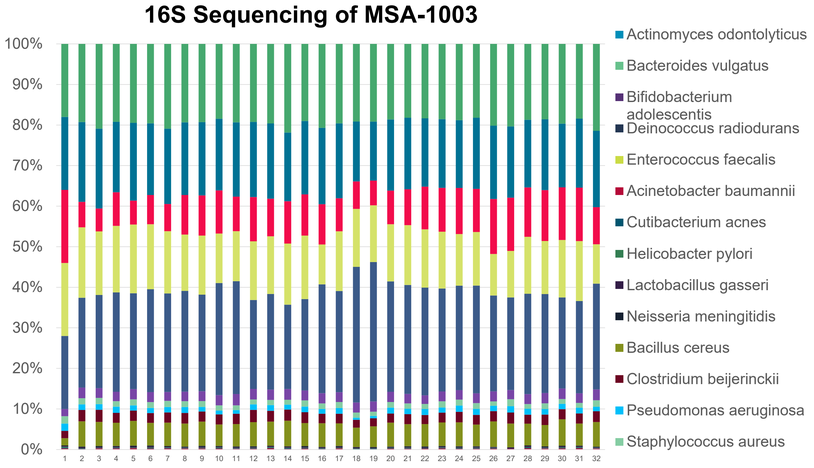

Figure 3. Replicates of MSA-1003 16S sequencing show high reproducibility.

HiFi Metagenomic Sequencing

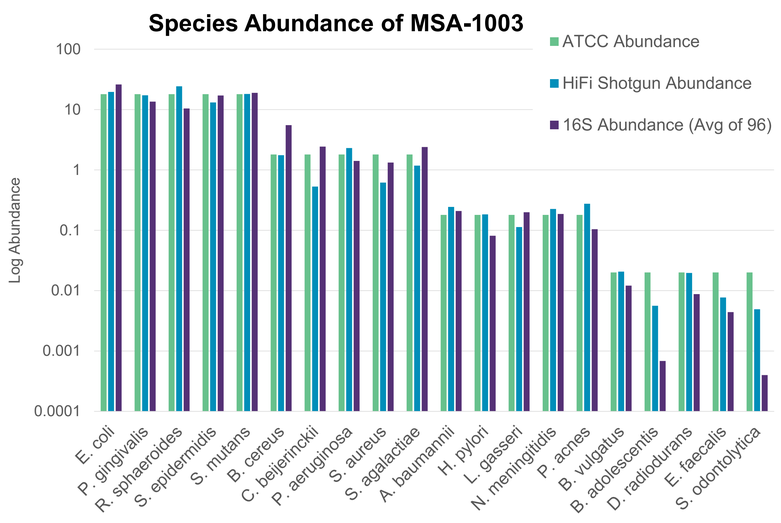

Figure 4. Shotgun sequencing of MSA-1003 (2b). Species down to 0.018 % abundance were detected successfully, and matches the average results seem across 96 replicates of16S sequencing of the same sample. Sequencing reads were mapped using BLAST. Download and explore the HiFi shotgun data set yourself.

Assembly-Free Gene Finding with FragGeneScan

Sample	# HiFi Reads	# amino acid sequences	genes / read	Mean protein size (amino acids)
Human fecal 1	2,802,471	16,429,903	5.9	320.2
Human fecal 2	1,646,208	11,993,089	7.3	325.1
Human fecal 3	1,593,641	13,054,811	8.2	321.7

Table 3. The long read length and high accuracy of HiFi reads means gene discovery can be done efficiently on unassembled metagenome sequences. As a result intact, error-free genes can be found even from species with too little coverage for assembly.

Metagenome Assembly and Binning

Figure 5. In cases where an assembly is desired, assembly of HiFi data with Canu outperforms alternative metagenome assembly approaches.
10X data: Bishara, A. et.al. (2018) High quality genome sequences of uncultured microbes by assembly of read clouds. Nature Biotechnology. 36, 1067-75.

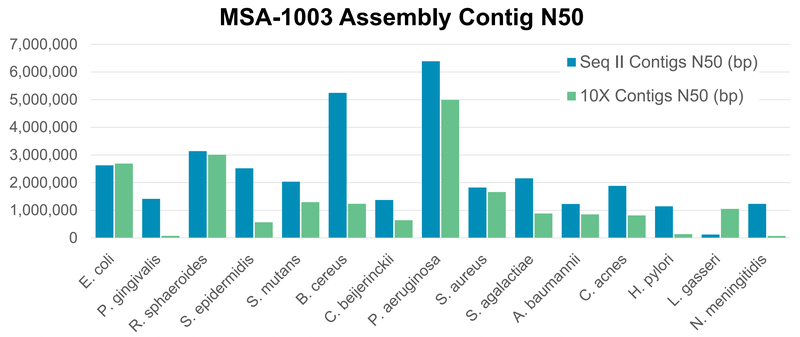

Table 4. Assembly of human fecal microbiome samples with Canu and binning with PATRIC /RAST Binning Service. RBS attempts to reconstruct complete genomes by seeding bins with one contig that encodes a unique 'seed role' protein, associating a reference genome with each bin, then using protein kmers to populate the bins. High quality bins have completeness >= 80%, fine consistency >= 87%, and contamination <= 10%.

Sample	# Contigs	# Bases	N50	RBS high quality bins	RBS total bins
Human fecal 1	7,043	370,834,836	221,615	11	64
Human fecal 2	11,993	583,831,665	123,296	21	118
Human fecal 3	4,945	275,125,396	182,086	10	61

Table 5. High quality RAST binning results for human fecal sample 2.

Genome	Score	Coarse consistency (%)	Fine consistency (%)	Completeness (%)
1	2020	97.9	97.3	100
2	2010	99.9	98.7	100
3	1973	97.5	96.2	100
4	1962	99.8	98.4	100
5	1818	99.2	98	96
6	1798	95.9	92.9	95.6
7	1684	99.3	97.9	99.7
8	1567	97.7	94.2	99.1
9	1547	95.6	93.6	90.2
10	1521	98.4	95	96.6
11	1505	96.8	94.4	96.7

Conclusions

- There is high correspondence between 16S and shotgun profiling data and expected mock community compositions, reflecting low context bias of SMRT sequencing technology.
- HiFi shotgun profiling enables the economical recovery of intact genes, operons, and predicted proteins, without the need for assembly.
- HiFi data can be analyzed with standard bioinformatic tools without modification.
- HiFi sequencing on the Sequel II platform provides a new option for functional profiling of microbiome samples.