Using the PacBio Sequel System to Taxonomically and Functionally Classify Metagenomic Samples in a Trial of Patients Undergoing Fecal Microbiota Transplantation

Richard Hall[1], Cheryl Heiner[1], Steve Oh[1], Chris Staley[2], Matt Hamilton[2], Alexander Khoruts[3] and Michael Sadowsky[2]
[1]PacBio, Menlo Park, CA
[2]BioTechnology Institute, University of Minnesota, St. Paul, MN
[3]Department of Medicine, University of Minnesota, Minneapolis, MN

Abstract

Whole-sample shotgun sequencing can provide a more detailed view of a metagenomic community than 16S sequencing, but its use in multi-sample experiments is limited by throughput, cost, and analysis complexity. While short-read sequencing technologies offer higher throughput, read lengths fewer than 500 bp rarely cover a gene of interest and necessitate assembly before further analysis. Assembling fragments requires sampling each community member at a high depth, significantly increasing the amount of sequencing needed and limiting the analysis of rare community members. Assembly methods also risk incorrectly combining sequences from different community members.

Single Molecule, Real-Time (SMRT) Sequencing reads in the 1-3 kb range, with >99% circular-consensus accuracy can be efficiently generated using the PacBio Sequel System. While base pair throughput is lower than some short-read technologies, the information content in 1-3 kb highly accurate reads are significantly higher than short-reads sequences. A high percentage of these long, highly accurate reads include gene fragements which can be used for analysis without the need for *de novo* assembly. As no assembly is required, the reads represent a random sampling of the community without requiring high coverage for each individual member. If one 2 kb read in a sampling of a hundred thousand can be uniquely identified by mapping to a database across its full length at high accuracy, we have high confidence in the presence of that microorganism even though its actual abundance could be as low as 0.00001%.

Samples from a blinded placebo-controlled trial have been sequenced on the Sequel System, including samples from pre-transplantation, post-heterologous FMT and post-autologous FMT. We demonstrate that the long-read metagenomic profiling workflow allows a high-resolution analysis of the differences between microbial communities, at the level of both taxonomic abundance and, by using gene prediction, function.

Long-Read Metagenomic Profiling Workflow

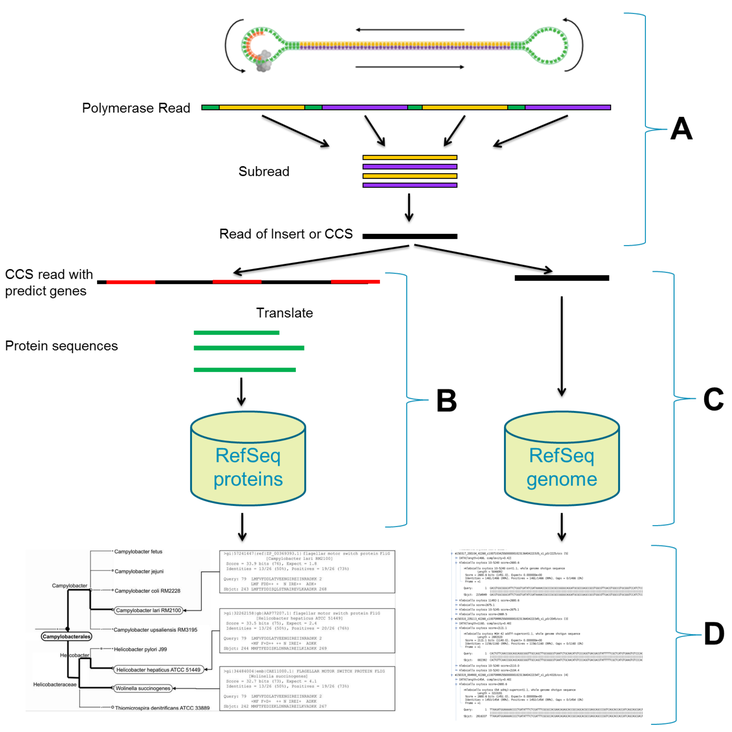


Figure 1. Analysis workflow for long-read metagenomic profiling

(A) Sheared genomic DNA with a mean length of ~2 kb is prepped and sequenced on the PacBio System. Multiple sequencing passes are made of the SMRTbell template, allowing the generation of high-quality circular consensus sequence (CCS) reads. **(B)** Prodigal (**Pro**karyotic **D**ynamic **P**rogramming **G**enefinding **Al**gorithm)[1] is used to predict genes in the consensus sequence and the amino acid sequence are calculated; blastp is used to align the putative protein sequences to the RefSeq bacterial protein database. **(C)** blastn is used to align the accurate CCS reads to the RefSeq genomic database. **(D)** Blast results from either method are imported into MEGAN[2] and a Lowest Common Ancestor (LCA) algorithm is used to assign a taxonomy to each sequence.

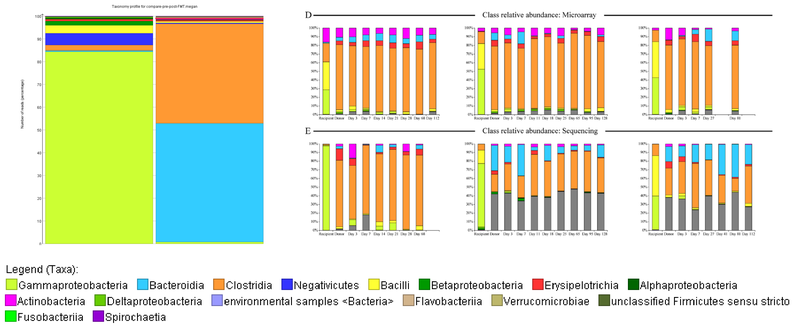
Sequencing

Sample	SMRT Cells	PacBio System	CCS (3 pass)	CCS N50	Predicted Genes	Genes / Read	Full-length Genes (Start Site, Stop Codon, RBS)	Full-length Genes / Read
Pre-FMT	12	PacBio RS II	80,299	1,373	177,006	2.42	76,868	1.05
Post-FMT	45	PacBio RS II	960,676	739	1,464,752	1.62	248,066	0.27
0 Week Pre-FMT	1	Sequel System Chemistry V1.0	76,996	1,952	249,410	3.24	153,530	2.00
2 Week A-FMT	1	Sequel System Chemistry V1.0	113,489	2,471	420,165	3.70	283,893	2.50
8 Week A-FMT	1	Sequel System Chemistry V1.0	59,674	2,349	204,357	3.42	138,051	2.31
2 Week H-FMT	1	Sequel System Chemistry V1.0	78,457	1,803	241,752	3.08	145,067	1.85
8 Week H-FMT	1	Sequel System Chemistry V1.0	75,124	2,282	274,956	3.66	178,428	2.38

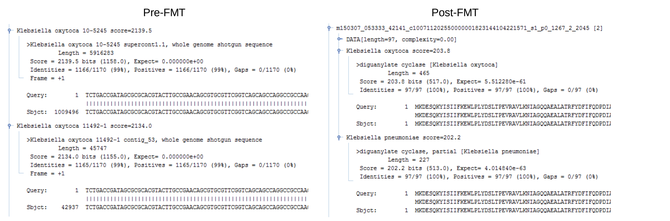
Table 1. Throughput from multiple FMT microbiome samples sequenced on the PacBio RS II or the Sequel System. The Pre- and Post-FMT samples were sequenced on the previous generation PacBio RS II system and are the samples discussed further in this poster. Samples FMT 3, 5 and 6 are similar, but distinct FMT samples run on the higher throughput Sequel System runs for comparison. Note the stats are affected not only by the sequencing system, but also by the library quality. A longer size distribution for the sequencing library will yield more predicted genes. The Sequel System and a library with a size distribution ~2 kb can yield >400,000 genes, with >250,000 being full length, as predicted by Prodigal (**Pro**karyotic **D**ynamic **P**rogramming **G**enefinding **Al**gorithm)[1].

FMT-Taxonomic and Functional Profile

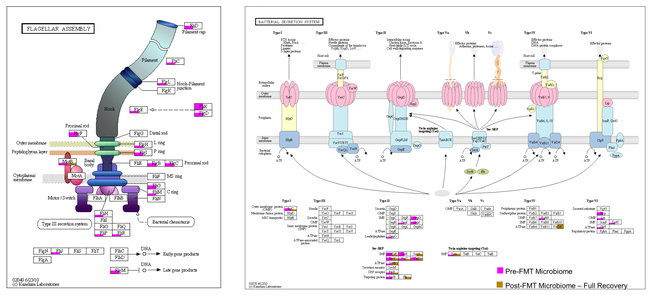
Figure 2. Taxonomic and functional profile of pre- and post-FMT samples from an individual suffering from chronic *C. difficile* infection.



(A) Class-level comparison with data from Microarray and 16S analysis[3,4]. The CCS method is demonstrated on a single individual, the published microarray and 16S data covers multiple individuals at different time points.



B) Example blast hits for both nucleotide and amino acid searches. In this case, the nucleotide classification has more power as the protein sequence is conserved across different species.



C) Each panel shows a KEGG[5] pathway and the frequency of homologous proteins found in the pre- and post-FMT samples. Proteins associated with flagella assembly are specific to the Pre-FMT sample. Bacterial Secretion System, type II and type VI secretion pathways are enriched in the Pre-FMT sample while the general Sec-SRP is found in both samples.

Comparison with Short-Read 16S, V5-V6 Region

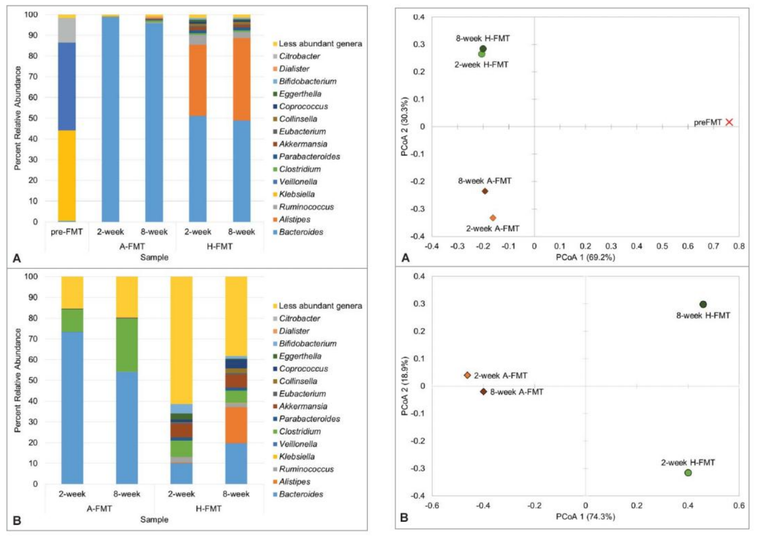


Figure 3. Distribution and principle coordinate analysis of genera in samples characterized using PacBio Sequel System (A) and Illumina 16S V5-V6 (B) platforms[6]. A-FMT: patients received autologous FMT; H-FMT: patients received heterologous FMT. Significant differences can be seen in the predicted distribution of genera between the two techniques. A greater relative abundances of *Bacteroides* and *Alistipes* is observed in samples characterized by PacBio, while those characterized by Illumina showed greater abundances of families found at lower abundances.

Principle coordinate analysis revealed similar separation of autologous FMT and heterologous FMT samples between platforms. However, post-heterologous-FMT samples analyzed using PacBio were considerably more alike than when analyzed using Illumina.

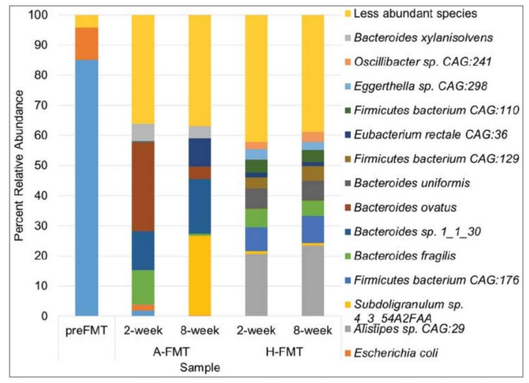


Figure 4. Species composition of samples characterized using the PacBio Sequel platform. A-FMT: patients received autologous FMT; H-FMT: patients received heterologous FMT. Resolution of communities at the species level revealed a greater shift in community composition among autologous FMT samples than heterologous FMT samples, autologous FMT samples are characterized by fluctuations in abundance of species predominantly within the genus *Bacteroides*. In contrast, heterologous FMT communities appeared more taxonomically stable and are comprised of a highly abundant species of *Alistipes* and more consistent distribution of *Bacteroides* spp.

Conclusion

- Long-read metagenomic profiling using CCS offers a unique data type that has distinct advantages over both 16S and shotgun assembly methods.
- High tolerance for sample input problems such as low input quantities and fragmented DNA.
- Allows species-level and, in some cases, strain-level taxonomic classification and functional studies.
- Sequel System throughput (chemistry V1.0) can yield > 100,00 reads and >145,000 full-length genes from a metagenomics community.

References

[1] Hyatt D, et al. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*. 28(17), 2223-2230.
[2] Huson DH, et al. (2011) Integrative analysis of environmental sequences using MEGAN 4, *Genome Research*. 2011. 21(9),1552-1560.
[3] Shankar V, et al. (2014) Species and genus level resolution analysis of gut microbiota in *Clostridium difficile* patients following fecal microbiota transplantation. *Microbiome*. 21(2), 13
[4] Quast C, et al., (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. 41,D590-596.
[5] Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 28(1), 27-30.
[6] Sadowsky M, et al. (2017) Analysis of gut microbiota – An ever changing landscape. *Gut Microbes*. doi:10.1080/19490976.2016.1277313