Workflow for Processing High-Throughput, Single Molecule, Real-Time Sequencing Data for Analyzing the Microbiome 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

Introduction

There are many sequencing-based approaches to understanding complex metagenomic communities spanning targeted amplification to whole-sample shotgun sequencing. While targeted approaches provide valuable data at low sequencing depth, they are limited by primer design and PCR. Whole-sample shotgun experiments generally use short-read sequencing, which results in data processing difficulties. For example, reads less than 500 bp in length will rarely cover a complete gene or region of interest, and will require assembly. This not only introduces the possibility of incorrectly combining sequence from different community members, it requires a high depth of coverage. As such, rare community members may not be represented in the resulting assembly.

Circular-consensus, Single Molecule, Real-Time (SMRT) Sequencing reads in the 1-3 kb range, with >99% accuracy can be generated using the previous generation PacBio RS II or, in much higher throughput, using the new Sequel System. While throughput is lower compared to short-read sequencing methods, the reads are a true random sampling of the underlying community since SMRT Sequencing has been shown to have very low sequence-context bias. With single-molecule reads >1 kb at >99% consensus accuracy, it is reasonable to expect a high percentage of reads to include genes or gene fragments useful for analysis without the need for *de novo* assembly.

Here we present the results of circular consensus sequencing for an individual's microbiome, before and after undergoing fecal microbiota transplantation (FMT) in order to treat a chronic *Clostridium difficile* infection. We show that even with relatively low sequencing depth, the long-read, assembly-free, random sampling allows us to profile low abundance community members at the species level. We also show that using shotgun sampling with long reads allows a level of functional insight not possible with classic targeted 16S, or short read sequencing, due to entire genes being covered in single reads.

Long-read Metagenomic Profiling Workflow

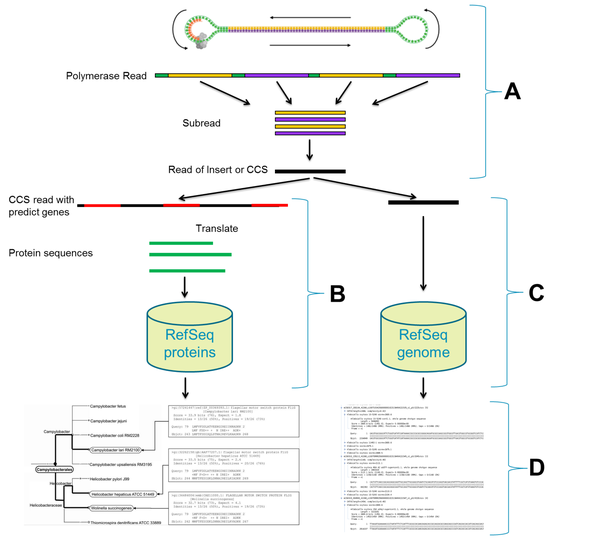


Figure 1. Analysis workflow for long-read metagenomic profiling

(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 used to align the putative protein sequences to the RefSeq bacterial protein database. **(C)** blastn 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.

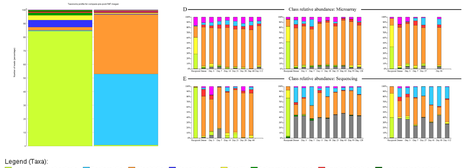
Sequencing Throughput

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
FMT 3	1	Sequel System	113,489	2,471	420,165	3.70	283,893	2.50
FMT 5	1	Sequel System	78,457	1,803	241,752	3.08	145,067	1.85
FMT 6	1	Sequel System	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 effected 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 new 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].

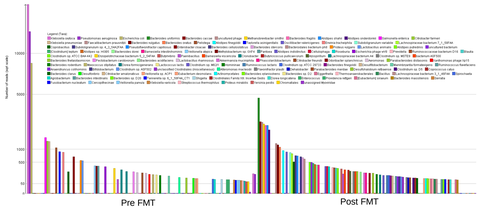
FMT - Taxonomic Profile

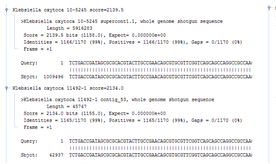
Figure 2. Taxonomic profile of pre- and post-FMT samples from an individual suffering from chronic *C. difficile* infection.

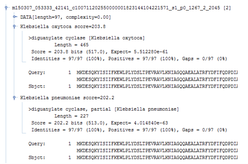


(A) Class-level comparison with data from Microarray and 16S analysis.[3] The CCS method is demonstrated on a single individual, the published microarray and 16S data covers multiple individuals at different time points.

(B) A high-resolution comparison of pre- and post-FMT samples. The long, highly accurate reads allow profiling at the species and in some cases the strain level.







(C) 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.

FMT – KEGG Functional Analysis

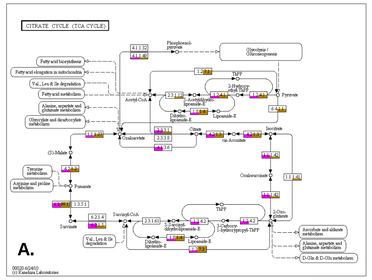
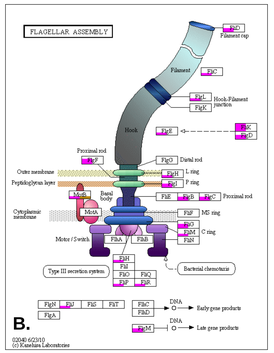
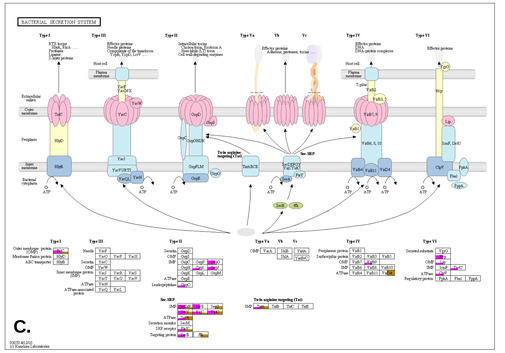


Figure 3. Gene prediction from long, accurate reads allows the identification of functional proteins

Each panel shows a KEGG[5] pathway and the frequency of homologous proteins found in the pre- and post-FMT samples. **(A)** The frequency of enzymes in the TCA pathway is consistent between both samples. **(B)** Proteins associated with flagella assembly are specific to the Pre-FMT sample. **(C)** 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.





Pre-FMT Microbiome

Post-FMT Microbiome – Full Recovery

Conclusion

Long-read metagenomic profiling using single-molecule CCS reads offers a unique data type that has distinct advantages over both 16S and shotgun assembly methods. While having a high tolerance for sample input problems such as low input quantities and fragmented DNA, long-read metagenomic profiling allows species-level and, in some cases, strain-level taxonomic classification and functional studies. Throughput for this kind of experiment on the Sequel System is such that a single sequencing run can yield > 145,000 full-length genes from a metagenomics community. For pre- and post-FMT microbiome samples, we show comparable results to both 16S and microarray data, while allowing finer-grain, species-level classification and functional insight.

References

[1] Hyatt D. et al., (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*. 28(17), 2223-2230.

[2] Huson D.H. 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.