Profiling the Microbiome in Fecal Microbiota Transplantation using Circular Consensus and Single Molecule, Real-Time Sequencing

Richard Hall[1], Cheryl Heiner[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

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 efficiently generated for low amounts of input DNA. 10 ng of input DNA sequenced in 4 SMRT Cells on the PacBio RS II would generate >100,000 such reads. 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 reads >1 kb at >99% accuracy it is reasonable to expect a high percentage of reads include 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.

Analysis Workflow

SMRT Sequencing and analysis workflow for 2 kb shotgun metagenomic library

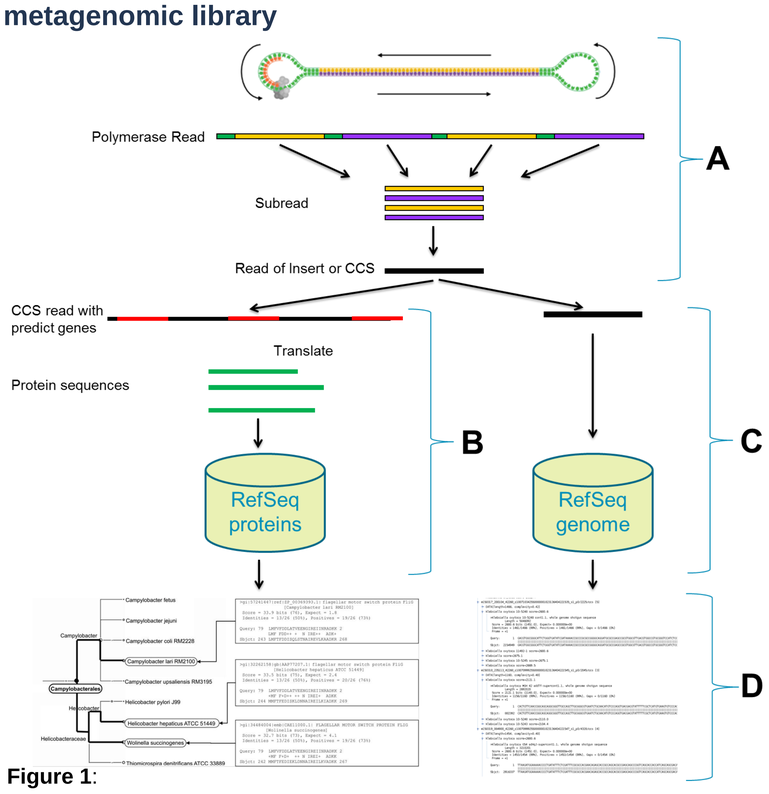


Figure 1:

A. Sheared genomic DNA with a mean length of ~2 kb is prepped and sequenced. Multiple sequencing passes are made of the SMRTbell template, allowing a high quality consensus to be generated.

B. Prodigal (**Pro**karyotic **D**ynamic **P**rogramming **G**enefinding **Al**gorithm)[1], is used to predict genes in the consensus sequence, 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 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.

Mock Microbiome

CCS – Gene Prediction vs. Full-Length 16S Mapping

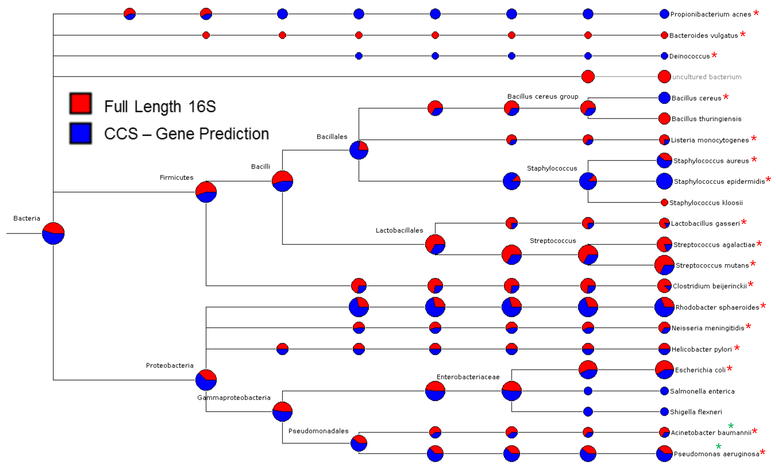


Figure 2: Shows the distribution of taxonomic hits, comparing the mapping of gene predictions from CCS reads to the mapping of full-length 16S amplicon reads to the Silva Database[3], both filtered at 0.0005% abundance. Pie charts on nodes show the total of all downstream assignments.

* Indicates known members of the mock community. * Incorrect assignment likely due to the sharing of genes between *Salmonella*, *Shigella* and *E. coli*.

FMT - Taxonomic Profile

Class-level comparison with data from Microarray and 16S analysis[4] (not same sample)

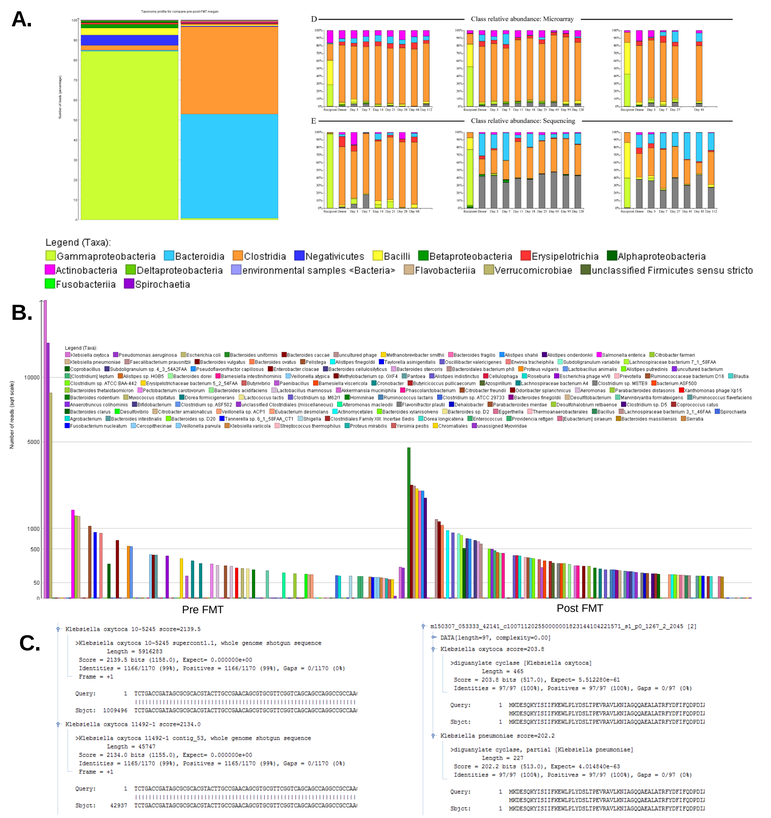


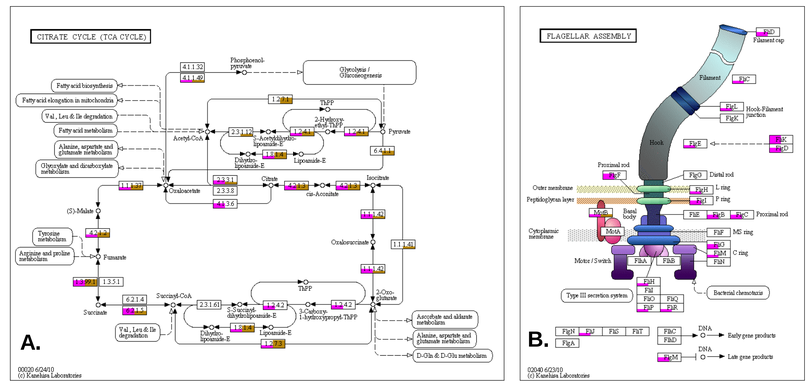
Figure 3: Taxonomic profile of an individual's fecal microbiome before FMT, while suffering from chronic *C. diff.* infection and post FMT.

A. Comparison of class level assignments between the proposed CCS method and published microarray and 16S analsysis. The CCS method is demonstrated on a single individual, the published microarray and 16S data covers multiple individuals at different time points.

B. High-resolution comparison of pre- and post-FMT samples, 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



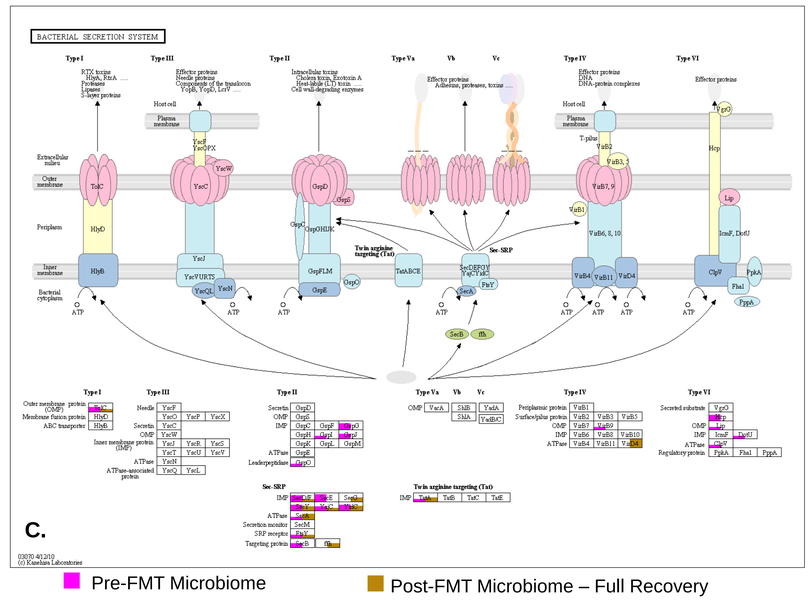


Figure 4: Prediction of genes in long, accurate reads allows the identification of functional proteins. Each panel shows a KEGG[5] pathway, and the frequency of homologous proteins found in the pre- and post-FMT samples.

A. TCA cycle, shown as a control, frequency of enzymes in the pathway is consistent between both samples.

B. Flagella assembly, 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.

Conclusion

Metagenomic profiling using single-molecule circular consensus 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, CCS profiling allows species-level, and in some cases strain-level, taxonomic classification and functional studies. Using a mock community, we show abundance measurements are comparable to 16S quantification, while assignment accuracy is higher. For real 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, Locascio P. F., Hauser L. J. and Uberbacher E.C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics.* 2012 Sep 1;28(17):2223-2230.

[2] Huson D.H. *et al*., Integrative analysis of environmental sequences using MEGAN 4, *Genome Res.* 2011. 21:1552-1560.

[3]Quast C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res*. 2013 41 (D1)

[4] Shankar V. *et al*. Species and genus level resolution analysis of gut microbiota in Clostridium difficile patients following fecal microbiota transplantation. *Microbiome.* 2014 Apr 21;2:13

[5] Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000 Jan 1;28(1):27-30.