

Profiling the Microbiome in Fecal Microbiota Transplantation using Circular Consensus and Single Molecule, Real-Time Sequencing Richard Hall¹, Cheryl Heiner¹, Chris Staley², Matt Hamilton², Alexander Khoruts³ and Michael Sadowsky²

¹PacBio, Menlo Park, CA, ²BioTechnology Institute, University of Minnesota, St. Paul, MN ³Department 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.

Mock Microbiome



FMT – KEGG Functional Analysis



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, assemblyfree, 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 shortread sequencing, due to entire genes being covered in single reads.

Analysis Workflow

SMRT Sequencing and analysis workflow for 2 kb shotgun

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 Database³, 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 analysis⁴ (not same sample)





metagenomic library



Gammaproteobacteria Bacteroidia Clostridia Negativicutes Bacilli Betaproteobacteria Erysipelotrichia Alphaproteobacteria Actinobacteria Deltaproteobacteria environmental samples <Bacteria> Flavobacteriia Verrucomicrobiae unclassified Firmicutes sensu stricto

Legend (Taxa): Klebsiella oxytoca Pseudomonas aeruginosa Escherichia coli Bacteroides uniformis Bacteroides caccae uncultured phage Methanobrevibacter smithii Bacteroides fragilis Alistipes shahii Alistipes onderdonkii Salmonella enterica Citrobacter farmeri Klebsiella pneumoniae Faecalibacterium prausnitzii Bacteroides vulgatus Bacteroides ovatus Pelistega Alistipes finegoldii Taylorella asinigenitalis Oscillibacter valericigenes Erwinia tracheiphila Subdoligranulum variabile Lachoospiraceae bacterium 7.1_58FAA Coprobacillus Subdoligranulum s.4_3_54A2FAA Pseudoflavonifractor capillosus Enterobacter cloacae Bacteroides celulosilyticus Bacteroides tervinas Cellulophaga Roseburia Escherichia phage w% Previetila Ruminicoccaceae bacterium D16 Blautia Clostridium sp. ATCC BAA-442 Erysipelotrichaceae bacterium 5.2_54FAA Butyrivibrio Paenibacillus Barnesiella viscericola Cronobacter Butyricicoccus pullicaecorum Azospirillum Lachnospiraceae bacterium ABCostridium sp. MSTE9 bacteroides distasonis Xanthomonas phage Xp15 Bacteroides rodentium Myxococcus stipitatus Orea formicigenerans Lactobacillus namnosus Akkermansia muciniphila Phascolarctobacterium Citrobacter feundii Odoribacter splanchnicus Aeromonas Parabacteroides distasonis Xanthomonas phage Xp15 Bacteroides rodentium Myxococcus stipitatus Dorea formicigenerans Lactobaccillus nsp. A8F502 unclassified Clostridium sp. A8F502 unclassified Clostridiales (miscellaneous) Alteromonas macleodii Flavonifractor plauti Dehalobacter Parabateroides finegoldii Desulfibohacterium Marvinbyantia sp. A5F502 unclassified Clostridiales (miscellaneous) Alteromonas macleodii Flavonifractor plauti Dehalobacter Parabateroides sp. D2 Tannerella sp. A5F1 Eubacterium desmolans Actinomycetales Bacteroides sp. D2 Tannerella sp. A5F1 Strata Spirochaeta Agrobacterium Nucleat



>Klebsiella oxytoca 10-5245 supercont1.1, whole genome shotgun sequence Length = 5916283 Score = 2139.5 bits (1158.0), Expect= 0.0000000e+00 Identities = 1166/1170 (99%), Positives = 1166/1170 (99%), Gaps = 0/1170 (0%) Frame = +1

ella oxytoca 11492-1 score=2134.0

Β.

>Klebsiella oxytoca 11492-1 contig_53, whole genome shotgun sequence Length = 45747 Score = 2134.0 bits (1155.0), Expect= 0.000000e+00 Identities = 1165/1170 (99%), Positives = 1165/1170 (99%), Gaps = 0/1170 (0%) DATA[length=97, complexity=0.00]
 Klebsiella oxytoca score=203.8

>diguanylate cyclase [Klebsiella oxytoca] Length = 465 Score = 203.8 bits (517.0), Expect= 5.512280e-61 Identities = 97/97 (100%), Positives = 97/97 (100%),

Query: 1 MKDESQKYISIIFKEWLPLYDSLTPEVRAVLKNIAGQQAEALATRFYDFIFQDPDL MKDESQKYISIIFKEWLPLYDSLTPEVRAVLKNIAGQQAEALATRFYDFIFQDPDL Sbict: 1 MKDESQKYISIFKEWLPLYDSLTPEVRAVLKNIAGQQAEALATRFYDFIFQDPDL

Y Klebsiella pneumoniae score=2

>diguanylate cyclase, partial [Klebsiella pneumoniae] Length = 227 Score = 202.2 bits (513.0), Expect= 4.014840e-63

O3070 4/12/10 (c) Kanehisa Laboratories Pre-FMT Microbiome

Post-FMT Microbiome – Full Recovery

Figure 4: Prediction of genes in long, accurate reads allows the identification of functional proteins. Each panel shows a KEGG⁵ 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.

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 MEGAN² and a Lowest Common Ancestor (LCA) algorithm is used to assign a taxonomy to each sequence.

Frame = +1		Identitie	s = 97/97 (100%), Positives = 97/97 (100%), Gaps = 0/97 (0%)
Query: 1	1 TCTGACCGATAGCGCGCACGTACTTGCCGAACAGCGTGCGT	Query:	1 MKDESQKYISIIFKEWLPLYDSLTPEVRAVLKNIAGQQAEALATRFYDFIFQDPDI;
			MKDESQKYISIIFKEWLPLYDSLTPEVRAVLKNIAGQQAEALATRFYDFIFQDPDI
Sbjct: 42937	7 TCTGACCGATAGCGCGCACGTACTTGCCGAACAGCGTGCGT	Sbict:	1 MKDESOKYISIIFKEWIPLYDSLTPEVRAVLKNIAGOOAEALATREYDFIFODPDI

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 analysis. 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.



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.

³Quast 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.

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2016 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. All other trademarks are the sole property of their respective owners.