Profiling Metagenomic Communities Using Circular Consensus and Single Molecule, Real-Time Sequencing

Richard Hall, Steve D. Oh, and Cheryl Heiner
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

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 amplification. Whole-sample shotgun experiments require 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 1-2 kb range, with >99% consensus accuracy can be efficiently generated for low amounts of input DNA, e.g. as little as 10 ng of input DNA sequenced in 4 SMRT Cells can generate >100,000 such reads. While throughput is low compared to second-generation sequencing, the reads are a true random sampling of the underlying community. Long read lengths translate to a high number of the reads harboring full genes or even full operons for downstream analysis.

Here we present the results of circular-consensus sequencing on a mock metagenomic community with an abundance range of multiple orders of magnitude, and compare the results with both 16S and shotgun assembly methods. We show that even with relatively low sequencing depth, the long-read, assembly-free, random sampling allows to elucidate meaningful information from the very low-abundance community members. For example, given the above low-input sequencing approach, a community member at 1/1,000 relative abundance would generate 100 1-2 kb sequence fragments having 99% consensus accuracy, with a high probability of containing a gene fragment useful for taxonomic classification or functional insight.

Low Input Sample Prep

Library Size	Input	# SMRT® Cells	Total Bases
10 kb (Standard Input)	1.25 µg	4	4.3 Gb
10 kb (Low Input)	125 ng	2	1.3 Gb
2 kb (Low Input)	10 ng	2	1.9 Gb

PACIFIC BIOSCIENCES®

Procedure & Checklist - Very Low (10 ng) Input 2 kb Template Preparation and Sequencing with Carrier DNA

Developed by Castle Raley, Leidos Biomedical Research, Inc. (formerly SAIC-Frederick)

For the full protocol, visit www.pacb.com/support/documentation

Microbiome Profiling

DNA was purified from an environmental (lake) sample and prepared for sequencing using the 2 kb, very low input Shared Protocol. Data was used to determine genes, which were aligned to a reference database. Resolution ranged from Phylum to Species with 15 Families, 6 Genus, and 3 Species found.

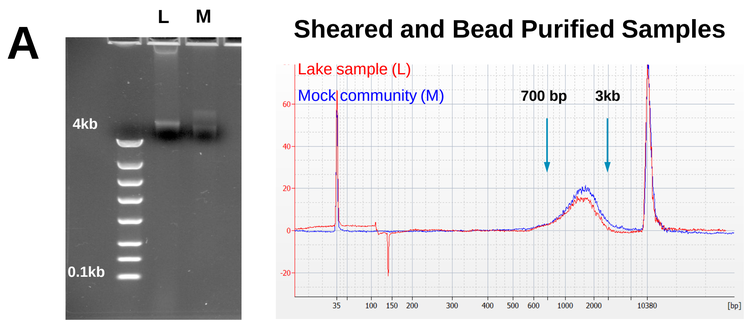

Figure 1. A. QC of microbiome samples prior to SMRTbell library prep **B.** Primary analysis of 2 kb sheared Lake Microbiome samples

Analysis Workflow

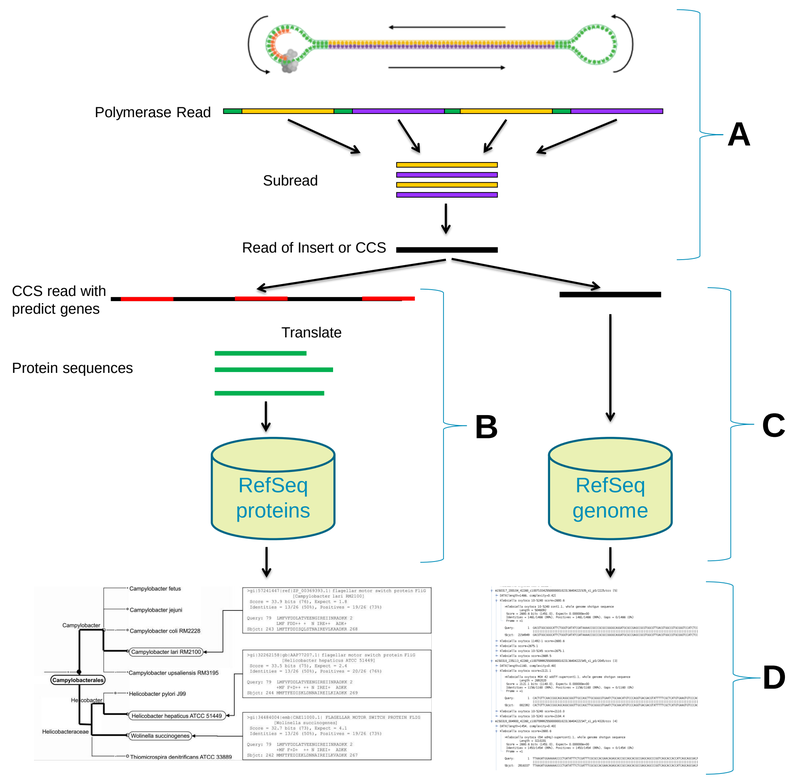

Figure 2. Workflow for analysis of a 2 kb shotgun metagenomic library. **A**. Sheared genomic DNA with a mean length of ~2 kb is prepped and sequenced. Multiple sequencing passes are made of the SMRTbell™ template, allowing a high quality circular consensus to be generated. **B**. Prodigal (**Pro**karyotic **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.

BEI Mock Metagenomic Community

CCS – Gene Prediction vs. Full Length 16S Mapping

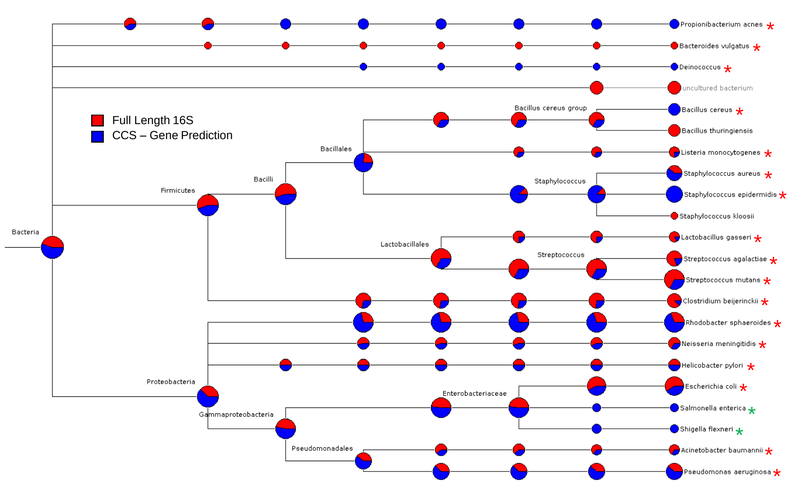

Figure 3, showing the distribution of taxonomic hits, comparing the mapping of gene predictions from CCS reads to the mapping of full length 16S amplicon reads to the Silva Database[4], both filtered at 0.0005% abundance. Pie charts on nodes show the total of all downstream assignments.
* Indicates known members of the mock community. * Incorrect assignment likely due to the sharing of genes between *Salmonella*, *Shigella* and *E. coli*.

CCS- Gene Prediction vs. Mapping Nucleotide Data

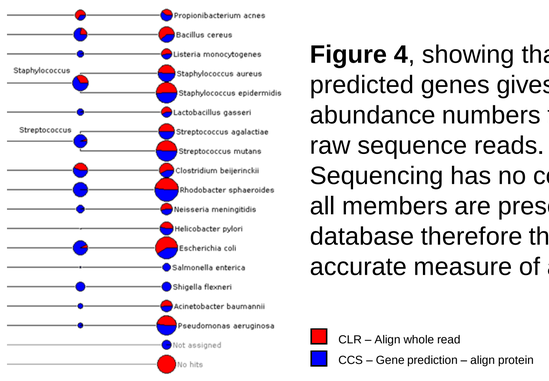

Figure 4, showing that mapping of predicted genes gives comparable abundance numbers to mapping of raw sequence reads. SMRT Sequencing has no context bias and all members are present in the database therefore this is an accurate measure of abundance.

CLR – Align whole read
CCS – Gene prediction – align protein

Lake Water Metagenome

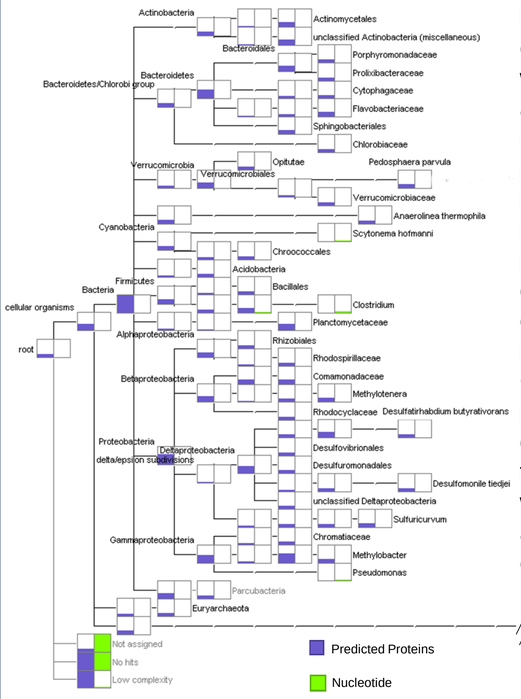

Figure 5. Taxonomic classification for a lake water sample. The analysis was carried out using both the gene prediction and the raw nucleotide methods. The nucleotide method produces very few hits, due to very few of the organisms being previously sequenced. Classification based on protein homology generates a rich taxonomic tree, and would be the suggested analysis workflow for environmental samples.

Human Gut Microbiome

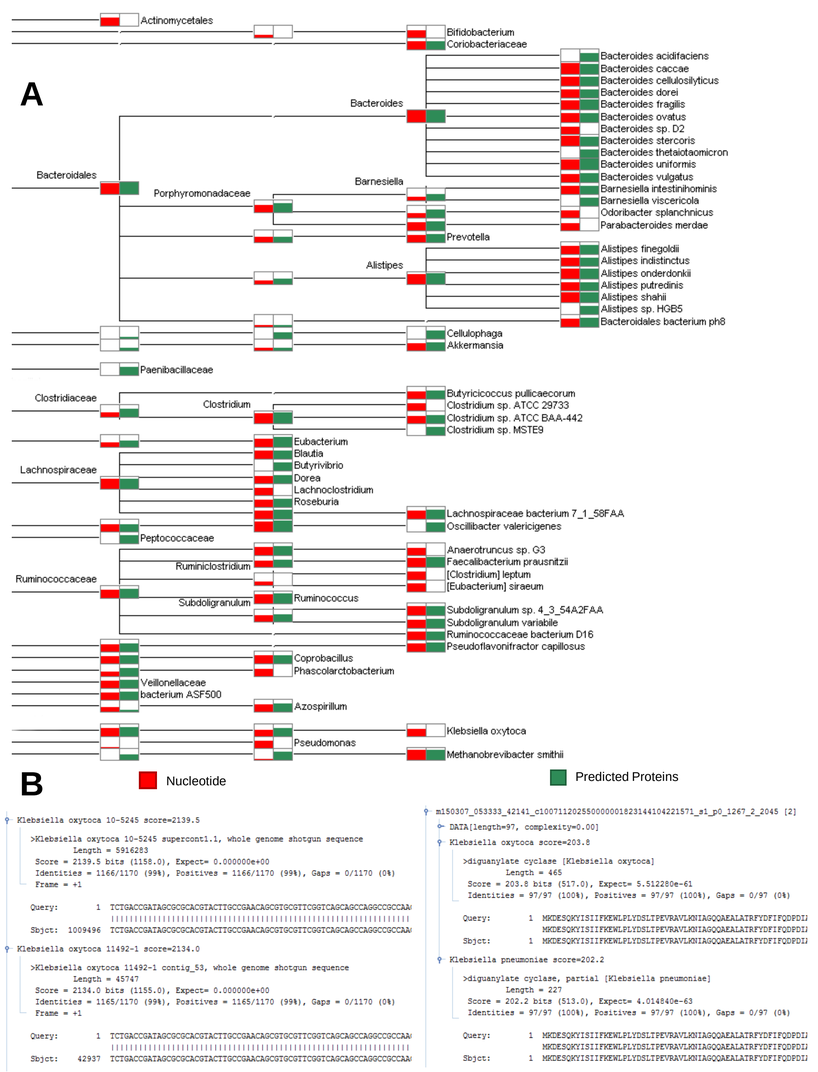

Figure 6. Taxonomic classification of human gut microbiome sample. **A,** taxonomic tree showing both protein and nucleotide based analysis. Unlike the environmental sample the classifications are comparable, due to better representation in the database. **B**, examples of specific alignments, in this case the nucleotide classification has more power as the protein sequence is conserved across different species.

References

References

[1] Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. 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] 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

[4] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Opens external link in new windowNucl. Acids Res. 41 (D1): D590-D596.

[5] Overbeek R, *et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005 Oct 7;33(17):5691-702*

Acknowledgements

Tanja Woyke, the Microbial Genomics Program Lead at the DOE Joint Genome Institute, for the lake, plant and mock metagenomic samples.
Michael Sadowsky, BioTechnology Institute, University of Minnesota, for the human microbiome samples