Developments in PacBio® Metagenome Sequencing: Shotgun Whole Genomes and Full-Length 16S

Chee How Lee[1], Brett Bowman[2], Cheryl Heiner[2], Richard Hall[2], Meredith Ashby[2], Jenny Gu[2] and Robert Tapella[2]

(1) Pacific Biosciences, Singapore; (2) Pacific Biosciences, Menlo Park, CA, United States

Abstract

The assembly of metagenomes is dramatically improved by the long read lengths of SMRT® Sequencing. This is demonstrated in an experimental design to sequence a mock community from the Human Microbiome Project, and assemble the data using the hierarchical genome assembly process (HGAP) at Pacific Biosciences. Results of this analysis are promising, and display much improved contiguity in the assembly of the mock community as compared to publicly available short-read data sets and assemblies. Additionally, the use of base modification information to make further associations between contigs provides additional data to improve assemblies, and to distinguish between members within a microbial community. The epigenetic approach is a novel validation method unique to SMRT Sequencing. In addition to whole-genome shotgun sequencing, SMRT Sequencing also offers improved classification resolution and reliability of metagenomic and microbiome samples by the full-length sequencing of 16S rRNA (~1500 bases long). Microbial communities can be detected at the species level in some cases, rather than being limited to the genus taxonomic classification as constrained by short-read technologies. The performance of SMRT Sequencing for these metagenomic samples achieved >99% predicted concordance to reference sequences in cecum, soil, water, and mock control investigations for bacterial 16S. Community samples are estimated to contain from 2.3 and up to 15 times as many species with abundance levels as low as 0.05% compared to the identification of phyla groups.

Datasets

All datasets are available for download:
https://github.com/PacificBiosciences/DevNet/wiki/Datasets

Human Microbiome Mock Community

Mock Community sample was obtained through BEI Resources, NIAID, and NIH as part of the Human Microbiome Project :

HM-276D Mixed bacteria Genomic DNA from Microbial Mock Community B (Even, High Concentration), v5.1H, for Whole Genome Shotgun Sequencing
(http://www.beiresources.org/Catalog/otherProducts/HM-276D.aspx)

The PacBio sequenced dataset consists of 49 SMRT Cells:

27 SMRT Cells using P4-C2 Chemistry
22 SMRT Cells using P5-C3 Chemistry

Full-length 16S amplicons

(in collaboration with Chunlab)

Several barcoded samples of full-length 16S amplicon sequences from three samples was obtained:

Water: Surface sea water sample from Oido, Siheung, Gyeonggi-do, South Korea.

Soil: Grassland soil from Seoul National University. Top soil (0-10 cm), pH 4.88, water content 28.5%

Cecum: Mouse cecum.

Full-length 16S Analysis

Full-length 16S sequencing coverage and accuracy:

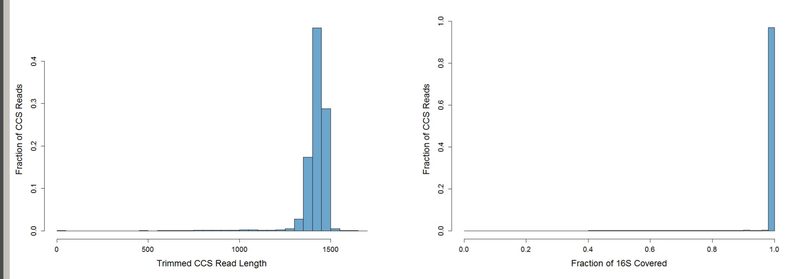


Figure 1. Distribution of trimmed CCS sequence lengths (left) and aligned CCS lengths relative to the canonical full-length 16S alignment (right).

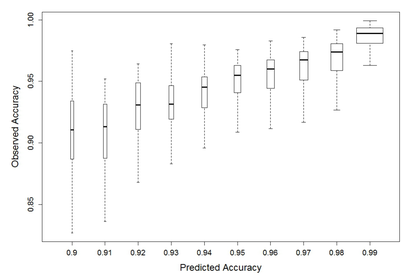


Figure 2. Predicted vs. observed accuracy for filtered CCS sequences. 80% of CCS reads fall into the two right-most boxes (>98%).

Full-length 16S analysis of several community samples:

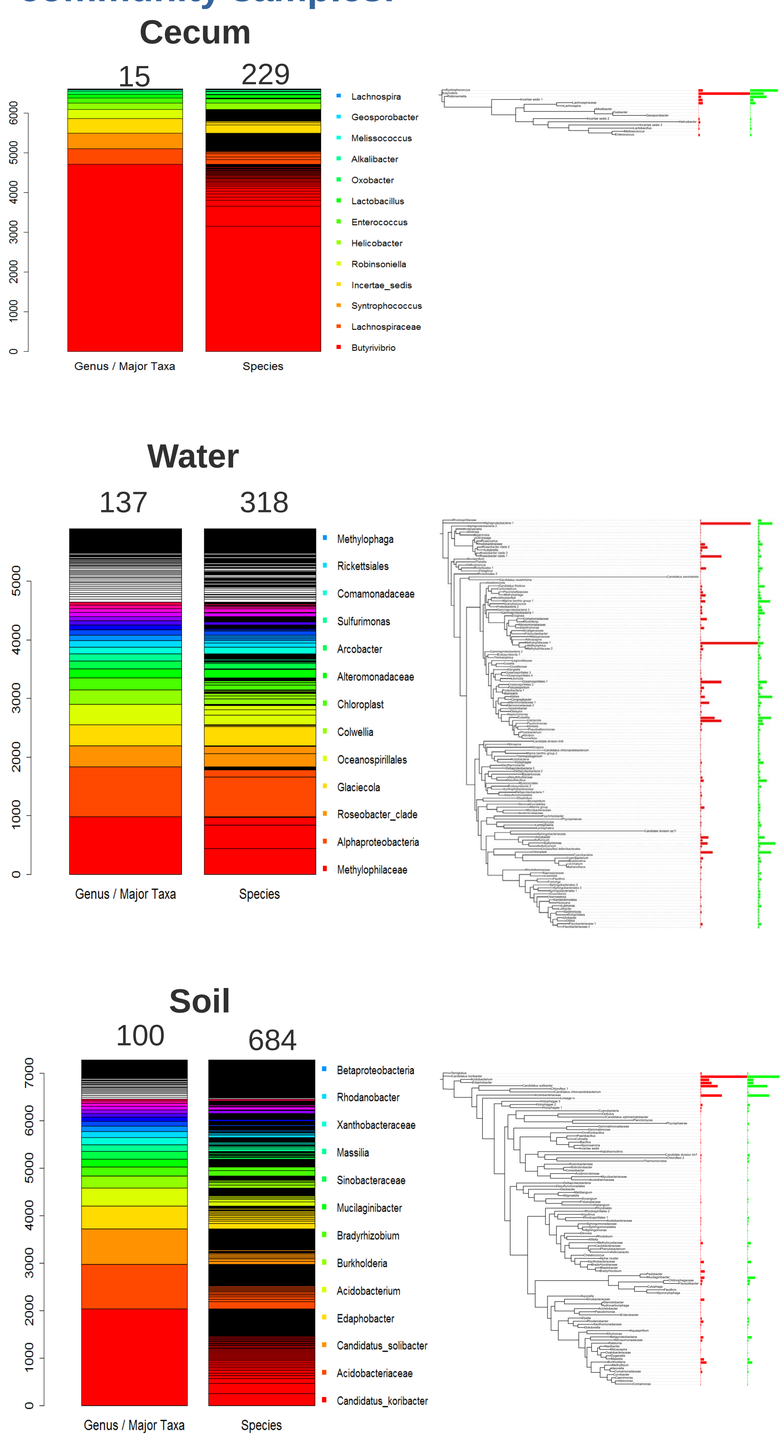


Figure 3. Compositional histograms of all metagenomic samples. The numbers above each bar denotes OTU count (Left). Phylogenetic trees [1] of all metagenomic samples by phylotype (Right). Bars denote the proportion of reads (Red) and species (Green) represented by each OUT. Differences between the bars suggest selection pressure. Sample data was analyzed using rDNA tools [2] and a custom pipeline that integrates the widely used Mothur suite[3] with PacBio-specific utilities.

Metagenomics Whole-genome Assembly

String graph assembly results for a mock human metagenome dataset:

37% assembled in **1 contig**
89% < 5 contigs

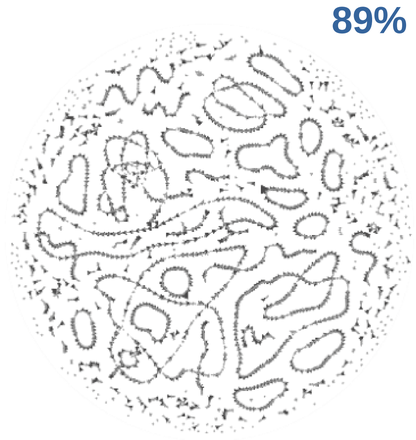
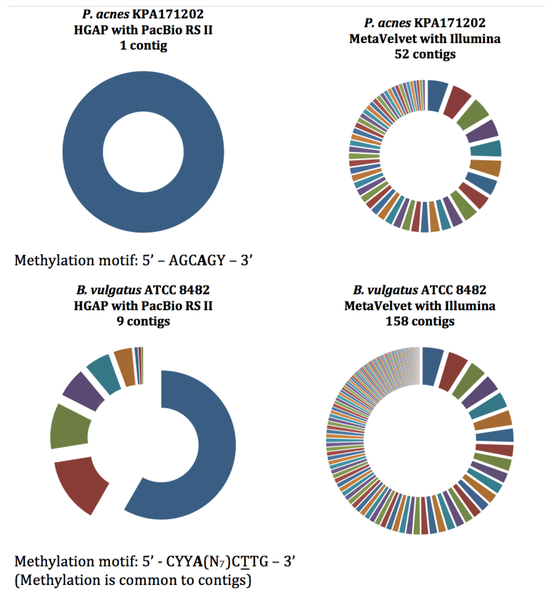


Figure 4. Schema visualization of string graph assembly results where circularity indicates closed, completed genomes of species within the population.10 microbes of near equi-molar concentrations were pooled.

Leveraging base-modification data to improve genome assemblies:



https://github.com/PacificBiosciences/DevNet/wiki/Human_Microbiome_Project_MockB_Shotgun

Figure 4. Comparison of whole-genome assembly results between PacBio long reads and competing short-read technology.

Conclusion

The PacBio® RS II provides a unique tool for sequencing full-length 16S to profile metagenomic communities and offers the potential to obtain complete genomes by using a whole-genome shotgun sequencing strategies. We offer:

- High-throughput classification to below the genus level
- Analysis of species richness **within** each OTU
- OTU consensus with 99.7-100% concordance to full-length reference
- Ability to assemble complete genomes down to one contig in metagenomic populations by additionally leveraging base modification data

References

[1] Letunic, I., & Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic acids research*, *39*(suppl 2), W475-W478.

[2] https://github.com/bnbowman/rDnaTools

[3] Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009. 75(23):7537-41

[4] Treangen and Koren et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline

ACKNOWLEDGEMENT:
We thank Mi Young Shin, Jong-Eun Lee, Yong-Joon Cho, and Jongsik Chun from DNALink and Chunlab for their contributions to the full-length 16s metagenomics project. We also thank Colleen Ludka for her assistance with the project.