# Analysis of Full-Length Metagenomic 16S Genes by SMRT® Sequencing

Brett Bowman [1], Mi Young Shin [2], Jong-Eun Lee [3], Yong-Joon Cho [4], Jongsik Chun [4]

[1] Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025; [2] Molecular Diagnostics Korea, Inc., Seocho-gu, Seoul, Korea; [3] DNA Link, Inc., Songpa-gu, Seoul, Korea; [4] Chunlab, Inc., Seoul National University, Seoul, Korea

## Introduction

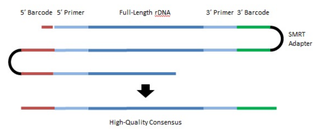
High-throughput sequencing of the 16S rRNA gene has become a valuable tool for characterizing microbial communities. However, the short read lengths produced by second-generation sequencing introduce bias and cannot provide taxonomic classification below the genus level, limiting biological insight. The best results are therefore obtained with full-length sequences of the 16S gene, however doing so with Sanger sequencing is time-consuming, costly, and low-throughput. SMRT® Sequencing has been used previously to analyze 16S amplicons[1], but to our knowledge no attempts have been made to analyze full-length (~1500 bp) 16S.

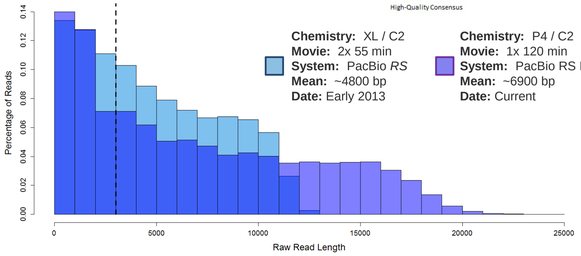
In this study, we evaluate the PacBio® *RS* and PacBio® RS II for full-length 16S rRNA gene sequencing and community profiling. We profiled four microbial communities (Cecum, Soil, Water, and a Mock control) for bacterial 16S, archaeal 16S, and fungal ITS using barcoded primers. Sample data was evaluated with rDnaTools[2], a custom pipeline that integrates the widely used Mothur suite[3] with PacBio-specific utilities.

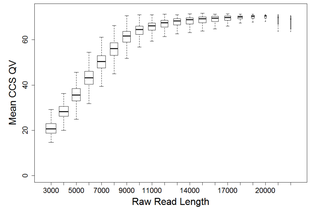
Each SMRT Cell of full-length 16S data generated 17-24k CCS reads, of which between 10-16k had >99% predicted concordance to the reference. Community samples were predicted to contain between 2.3-15 times as many species as phyla, with abundance levels as low as 0.05%. SMRT® Sequencing therefore represents a novel platform for community analysis, allowing for high depth and unprecedented sensitivity.

## Full-Length 16S Sequence

**Circular Consensus Sequences (CCS)** PacBio's long read length and circularized templates provide high-quality consensus from multiple passes over the same molecule




**Assaying 16S CCS**

*Above:* Raw read length distributions for sequencing of full-length 16S (~1500 bp) amplicons. The dotted vertical line denotes the minimum read length for CCS.

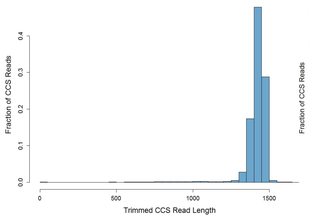
*Left:* Mean quality-values for full-length 16S CCS sequences by raw read length.

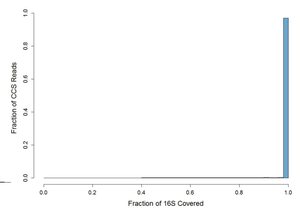
## Analysis Workflow

Analysis of 16S CCS sequences was carried out with standard tools from the Mothur package (http://www.mothur.org)[3] where possible. Where existing tools were not well suited for analyzing SMRT sequence data, custom scripts were written using Python. The complete analysis pipeline, and all custom scripts are available for download on GitHub (https://github.com/bnbowman/rDnaTools). Steps highlighted in **bold** are implemented with custom scripts.

1. **Export CCS Sequence Data from Bas.H5**
2. **Filter Sequences by Quality**
3. **Group Sequences by Barcode and/or Primer**
4. Align Sequences to Reference
5. Filter Partial Alignment Reads (<80%)
6. Identify & Remove Chimeric Sequences
7. Pre-Cluster Sequences
8. Calculate Distance Matrix
9. Cluster Sequences
10. **Generate OTU Consensus Sequences**
11. Classify Clusters based on the Consensus
12. Summarize OTUs

## Analysis Results



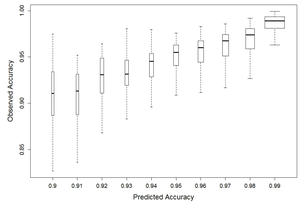




**Analyzing Full-Length 16S**

*Above-Left:* Distribution of trimmed CCS sequence lengths

*Above-Right:* Distribution of aligned CCS lengths relative to the canonical full-length 16S alignment

*Right:* Predicted versus observed accuracy for filtered CCS sequences. 80% of CCS reads fall into the two right-most boxes (>98% )

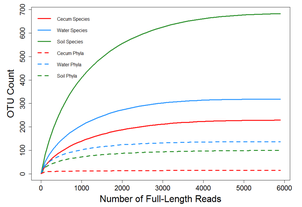


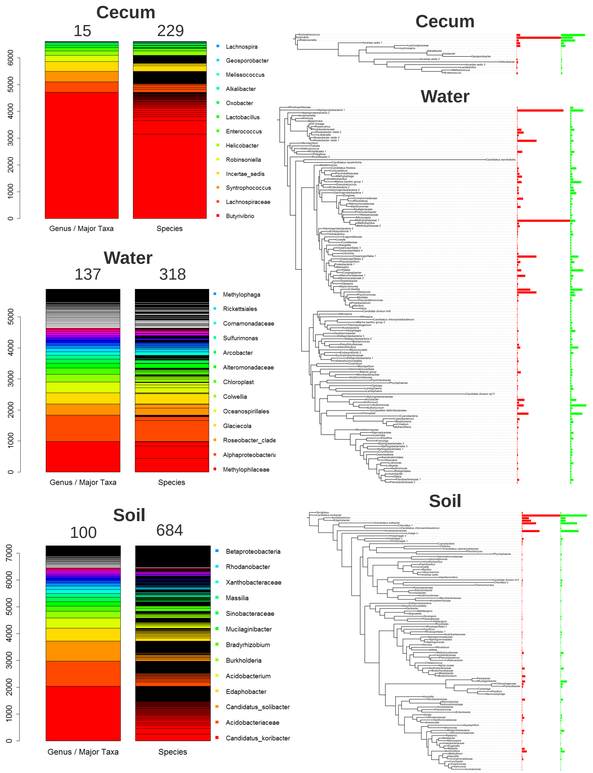

## Community Analysis

*Right*: Rarefaction curves for all metagenomic samples at both the genus and species level.

*Below-Left*: Compositional histograms of all metagenomic samples. The numbers above each bar denotes OTU count.

*Below-Right*: Phylogenetic trees[4] of all metagenomic samples by phylotype. Bars denote the proportion of reads (Red) and species (Green) represented by each OTU. Differences between the bars suggest selection pressure.

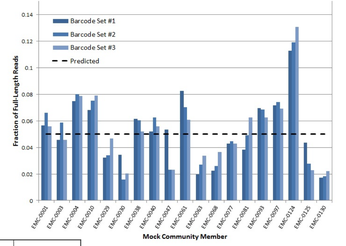



## Sample Multiplexing

**Barcoded 16S Samples**

- >99.5% of CCS sequences had at least 1 recognizable barcode
- All samples showed high reproducibility between barcoded replicates
- At 3x multiplexing, each SMRT cell generates ~5k full-length 16S sequences per sample with >99% predicted accuracy




| Amplicon | Total | >99% | >99.5% |
|---|---|---|---|
| 1.5 kbp | 23952 | 15136 | 10591 |
| 1 kbp* | 27095 | 19629 | 15417 |
| 750 bp* | 29136 | 22306 | 18209 |
| 500 bp* | 31035 | 25273 | 21544 |

**Alternative Amplicons**

CCS quality depends primarily on the number passes, allowing for trade-offs between quality, amplicon size, and throughput. For example, researchers interested in the fungal ITS (~700 bp) could choose to either raise their quality threshold or multiplex an additional sample relative to full-length 16S.

*These are predicted values

## Conclusions

The PacBio® RS II provides a unique tool for sequencing full-length 16S that can simultaneously profile a metagenomic community and generate high-quality, full-length reference sequences:

- High-throughput classification to below the genus level
- Analysis of species-richness **within** each OTU
- 17-24k high-quality CCS sequences per SMRT Cell (~2 hour run)
- 10-16k CCS sequences with >99% predicted concordance
- 96% of CCS reads covered the full-length of the canonical alignment
- OTU consensus with 99.7-100% concordance to full-length reference
- Flexible trade-offs between length, quality, and throughput
- Bioinformatics tools for sequence filtering, multiplexing, and consensus

## References

[1] Fichot, E.B., and Norman, R.S. "Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform." *Microbiome* 1.1 (2013): 1-5.

[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] 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.