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 II and PacBio® RS 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. Analysis of species-richness (Red) and species (Green) represented by each OTU. Differences between the bars suggest selection pressure. The PacBio® RS II provides a unique tool for sequencing full-length 16S that can anatomically 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)
- >99% CCS sequences with >99% predicted concordance
- 96% of CCS reads covered the full-length of the canonical alignment
- OTU consensus with 97.1% concordance to full-length reference
- Flexible trade-offs between length, quality, and throughput
- Bioinformatics tools for sequence filtering, multiplexing, and consensus analysis

Conclusions

The PacBio® RS II provides a unique tool for sequencing full-length 16S that can anatomically 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)
- >99% CCS sequences with >99% predicted concordance
- 96% of CCS reads covered the full-length of the canonical alignment
- OTU consensus with 97.1% concordance to full-length reference
- Flexible trade-offs between length, quality, and throughput
- Bioinformatics tools for sequence filtering, multiplexing, and consensus analysis

References


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

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 II and PacBio® RS 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.

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/boltonam/DnaTools). 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 (<60%)
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

Analysis of Full-Length 16S

Above-Left: Distribution of trimmed CCS sequences lengths by raw read length.

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

Right-Flattened curves for all metagenomic samples at both the genus and species level. Below-Left: Comparative histograms of all metagenomic samples. The numbers above each bar denote OTU count. Below-Right: Phylogenetic tree 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.

Community Analysis

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 16-24k full-length 16S sequences per sample with >95% predicted accuracy

Sample Multiplexing

- >99.5% accuracy
- CCS quality depends primarily on the number of 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 thresholds or multiplex an additional sample relative to full-length 16S.
- These are predicted values

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 CCS sequences by raw read length.

Analysis of Full-Length 16S

Above-Left: Distribution of trimmed CCS sequences lengths by raw read length.

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

Right-Flattened curves for all metagenomic samples at both the genus and species level. Below-Left: Comparative histograms of all metagenomic samples. The numbers above each bar denote OTU count. Below-Right: Phylogenetic tree 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.

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/boltonam/DnaTools). 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 (<60%)
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