Accurately Surveying Uncultured Microbial Species with SMRT® Sequencing

B. Bowman[1], J-F. Cheng[2], E. Singer[2], D. Coleman-Derr[2], S. Tringe[2], S. Hallam[3], T. Woyke[2]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, [2]Dept. of Energy Joint Genome Inst., Walnut Creek, CA, [3]Univ. of British Columbia, Vancouver, BC, Canada

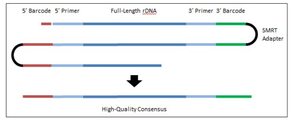
Introduction

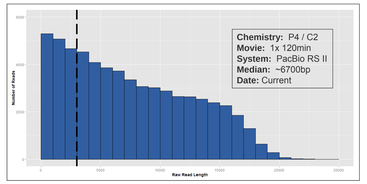
Microbial ecology is reshaping our understanding of the natural world by revealing the large phylogenetic and functional diversity of microbial life. However the vast majority of these microorganisms remain poorly understood, as most cultivated representatives belong to just four phylogenetic groups and more than half of all identified phyla remain uncultivated. Characterization of this microbial 'dark matter' will thus greatly benefit from new metagenomic methods for *in situ* analysis. For example, sensitive high throughput methods for the characterization of community composition and structure from the sequencing of conserved marker genes.

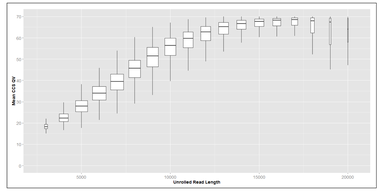
Here we utilize Single Molecule Real-Time (SMRT®) sequencing of full-length 16S rRNA amplicons to phylogenetically profile microbial communities to below the genus-level. We test this method with three different data sets: a mock community of known composition, a previously studied microbial community from a lake known to predominantly contain poorly characterized phyla[1], and a multiplexed. These results are compared to traditional 16S tag sequencing of the V4 hyper-variable region with short-read sequencing technologies. We explore the benefits of using full-length ribosomal DNA amplicons for estimating community structure and diversity, as well as for single-cell isolate strain identification relative to alternate methods. We characterize the potential benefits of profiling metagenomic communities with full-length 16S rRNA genes from SMRT sequencing relative to standard methods.

Reads-of-Insert Consensus

Reads-of-Insert (RoI) Consensus Sequences
PacBio's long read-length and circularized sequence templates enable the generation of high-quality consensus sequences from multiple passes over the same molecule.







Assaying 16S Reads-of-Insert Sequence Quality

Figure 1 (Right): A diagram of the generation of a Read-of-Insert consensus from an unrolled sequence consisting of 1 full and 2 partial pass "subreads"

Figure 2 (Top-Left): Raw read length distributions for sequencing of full-length 16S (~1500bp) amplicons. The dotted vertical line denote the minimum read length for the generation of Reads-of-Insert consensus.

Figure 3 (Bottom-Left): Mean quality-values for full-length 16S CCS sequences by raw (unrolled) sequencing read length. Box-width denotes relative abundance.

Improved Accuracy from Full-Length 16S

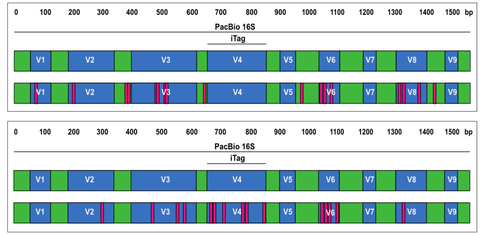


Figure 4.
A) Comparison of differences between two *Salmonella* spp. in the full-length 16S gene (2.6%) and hyper-variable V4-region (0%)
B) Example of the over-estimation of diversity that can result from classification based on small hyper-variability regions compared to the full-length sequence

Environmental Sample

Community analysis of environmental 16S sequence data from Lake Sakinaw, British Columbia. Illumina V4-region data was analyzed with iTagger[2], while PacBio full-length 16S data was analyzed with rDnaTools[3].

Figure 5 (Right): Phylogenetic tree of all OTUs (359 assigned) identified from full-length PacBio 16S sequences[4]. ***Pink:*** 175 (48.74%) unclassified OTUs; ***Orange:*** candidate phyla; ***Blue:*** phyla classified in PacBio data and absent from iTags. Branch colors denote phyla clades.

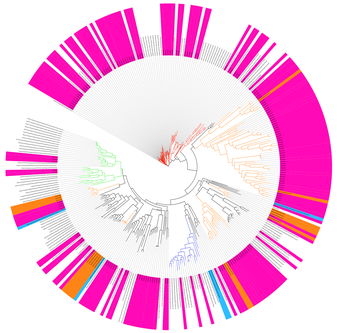
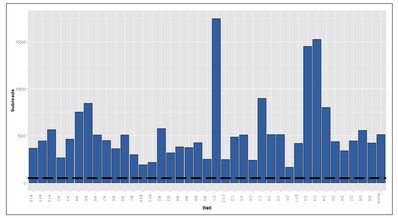
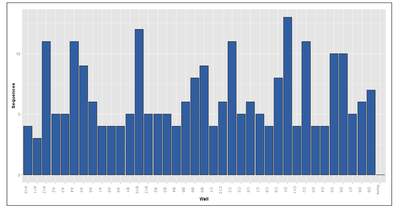


Table 1 (Below): Counts of clusters and reads at various levels of assignment for Illumina iTag and PacBio 16S data. Data sets show strong concordance with each other except at the OTU level, where the

Grouping	iTags	PacBio 16S
# of phyla resolved (reads)	32 (4,407)	34 (5,000)
# of candidate phyla (reads)	18 (2,139)	16 (1,724)
# of families resolved (reads)	45 (3,906)	49 (4,636)
# of families from candidate phyla (reads)	22 (3,197)	18 (3,217)
# of OTUs (reads)	1,843 (4,407)	359 (5,000)

Multiplexed Strain Identification





Pilot project for multiplexed strain identification from barcoded full-length 16S sequence data using asymmetric, paired barcodes and Long Amplicon Analysis.

Figure 6 (Left-Top):
Histogram of post-filter subread counts by barcode. Unlabeled subreads account for < 2.5% of the over 200,000 high-quality subreads. The dashed line represents the 500 subreads recommended for phasing.

Figure 7 (Left-Bottom):
Histogram of consensus sequence counts by barcode.
Multiple consensus sequences enable better classification,

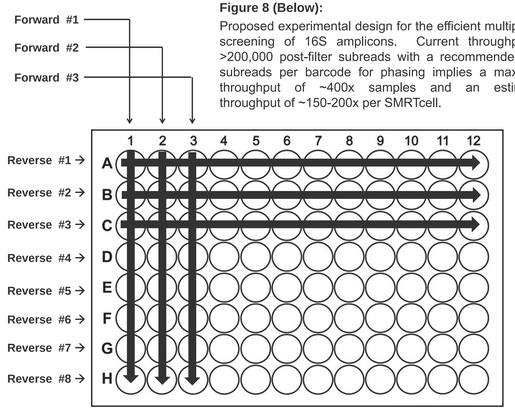


Figure 8 (Below):
Proposed experimental design for the efficient multiplexed screening of 16S amplicons. Current throughput of >200,000 post-filter subreads with a recommended 500 subreads per barcode for phasing implies a maximum throughput of ~400x samples and an estimated throughput of ~150-200x per SMRTcell.

Mock Community Classification

Species	OTU Count	Concordance w/ Reference	V4 Class.	Full Class.
Clostridium perfringens	1	99.92%	Genus	Species
Clostridium thermocellum	1	99.86%	Genus	Strain
Coraliomargarita akajimensis	1	100%	Genus	Strain
Corynebacterium glutamicum	1	99.79%	Genus	Strain
Desulfosporosinus acidiphilus	1	100%	Genus	Strain
Desulfosporosinus meridiei	1	99.67%	Genus	Strain
Desulfotomaculum gibsoniae	5	98.2%-100%	Family	Strain
Echinicola vietnamensis	1	100%	Genus	Strain
Escherichia coli	1	99.86%	Family	Species
Fervidobacterium pennivorans	1	99.93%	Genus	Strain
Frateuria aurantia	1	99.93%	Genus	Strain
Hirschia baltica	1	100%	Genus	Strain
Meiothermus silvanus	1	100%	Genus	Species
Olsenella uli	1	100%	Genus	Strain
Pseudomonas stutzeri	1	100%	Genus	Strain
Salmonella bongori	2	99.93%	Family	Strain
Salmonella enterica	1	99.24%	Genus	Serovar
Segniliparus rotundus	1	99.93%	Genus	Strain
Spirochaeta smaragdinae	1	100%	Genus	Strain
Streptococcus pyogenes	1	100%	Genus	Strain
Terriglobus roseus	1	100%	Genus	Strain
Thermobacillus composti	1	100%	Genus	Strain

Analysis of bacterial Mock Community, composed of a mixture of 22 bacterial species for which reference assemblies was available. Full-length 16S amplicons were generated with the 27F/1492R primer pair and sequenced on the PacBio RS II. Sequence data was analyzed with rDnaTools[3].

Table 2 (Left):
Analysis of the OTU consensus sequences generated by rDnaTools. Classification of the V4 region was carried out with the RDP Classifier[5] on V4 sequences extracted from the reference genomes. Classification of full-length 16S was performed with RDP SeqMatch[5] on the OTU consensus sequences generated by rDnaTools

Conclusion

- The PacBio RSII enables high-throughput sequencing of full-length 16S amplicons
- Full-length 16S amplicons show both greater sensitivity and specificity to the clustering of OTUs than 16S tag sequencing
- Comparisons of short-read and full-length 16S sequences from environmental samples show high concordance
- SMRT® Sequencing could be a robust, cost-effective method for multiplexed strain identification

References

[1] Rinke, Christian, et al. "Insights into the phylogeny and coding potential of microbial dark matter." *Nature* (2013).
[2] https://bitbucket.org/berkeleylab/jgi_itagger
[3] https://github.com/PacificBiosciences/rDnaTools
[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.
[5] Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis*Nucl. Acids Res.* 41(Database issue):D633-D642; doi: 10.1093/nar/gkt1244 [PMID: 24288368]