High-resolution phylogenetic microbial community profiling

Esther Singer[1*], Devin Coleman-Derr[1], Brett Bowman[2], Patrick Schwientek[1], Alicia Clum[1], Alex Copeland[1], Doina Ciobanu[1], Jan-Fang Cheng[1], Esther Gies[3], Steve Hallam[3], Susannah Tringe[1], Tanja Woyke[1]

[1] LBNL - Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

[2] Pacific Biosciences, 1380 Willow Road, Menlo Park, CA USA

[3] University of British Columbia, Vancouver, BC Canada

**To whom correspondence should be addressed*: Email: Esther Singer: esinger@lbl.gov or Devin Coleman-Derr: dacoleman-der@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER:

LBNL: This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

High-resolution phylogenetic microbial community profiling

Esther Singer[1], Devin Coleman-Derr[1], Brett Bowman[2], Patrick Schwientek[1], Alicia Clum[1], Alex Copeland[1],
Doina Ciobanu[1], Jan-Fang Cheng[1], Esther Gies[3], Steve Hallam[3], Susannah Tringe, [1] Tanja Woyke[1]

1: Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598; 2: Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025; 3: University of British Columbia, Vancouver, BC V6T 1Z3, Canada

Abstract

The representation of bacterial and archaeal genome sequences is strongly biased towards cultivated organisms, which belong to merely four phylogenetic groups. Functional information and inter-phylum level relationships are still largely underexplored for candidate phyla, which are often referred to as 'microbial dark matter'. Furthermore, a large portion of the 16S rRNA gene records in the GenBank database are labeled as "environmental samples" and "unclassified", which is in part due to low read accuracy, potential chimeric sequences produced during PCR amplifications and the low resolution of short amplicons. In order to improve the phylogenetic classification of novel species and advance our knowledge of the ecosystem function of uncultivated microorganisms, high-throughput full length 16S rRNA gene sequencing methodologies with reduced biases are needed. We evaluated the performance of PacBio single-molecule real-time (SMRT) sequencing in high-resolution phylogenetic microbial community profiling. For this purpose, we compared PacBio and Illumina metagenomic shotgun and 16S rRNA gene sequencing of a mock community as well as of an environmental sample from Sakinaw Lake, British Columbia. Sakinaw Lake is known to contain a large percentage of microbial species from candidate phyla. Sequencing results show that community structure based on PacBio shotgun and 16S rRNA gene sequences is highly similar in both the mock and the environmental communities. Resolution power and community representation accuracy from SMRT sequencing data appeared to be independent of %GC content of microbial genomes and was higher when compared to Illumina-based metagenome shotgun and 16S rRNA gene (iTag) sequences, e.g. full-length sequencing resolved all 23 OTUs in the mock community, while iTags did not resolve closely related species. SMRT sequencing hence offers various potential benefits when characterizing uncharted microbial communities.

Objectives

This study pursued the following aims:

1) To evaluate the representation accuracy between PacBio and Illumina (shotgun and 16S) of a known community
2) To evaluate the influence of GC-richness on sequence representation
3) To evaluate the resolution power of the phylogenetic diversity of Sakinaw Lake, especially with respect to candidate phyla
4) To evaluate the representation of candidate phyla OTUs and compare between PacBio 16S and iTags

Results

Mock Community Analysis

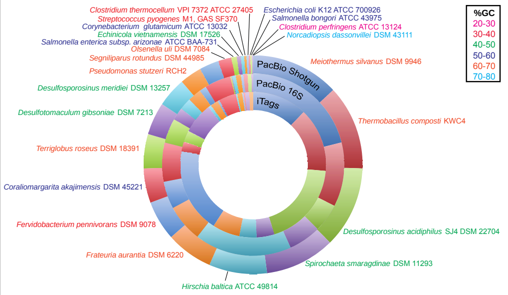

Figure 1: Abundance profiles of Mock community as represented by PacBio and Illumina sequences. Order follows the abundance profile of the PacBio Shotgun sequence data. Species name are colored by %GC range.

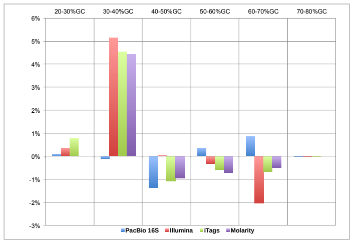

Figure 2: Average abundance differences according to %GC range. Species abundance as represented by PacBio Shotgun data was used as reference, to which all other datasets were compared.

Sakinaw Lake Analysis

Table 1: OTU-assignment statistics of classifiable sequences. Comparison between iTag and PacBio 16S data shows that datasets are very similar and do not show large discrepancies where sequences can be classified.

	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)

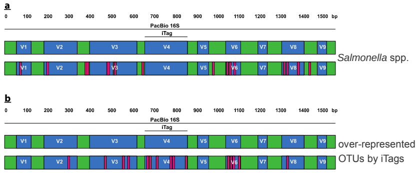

Figure 3: a) Sequence similarities between *Salmonella* spp. 16S full-length (97.4%) and V4-region (100%) sequences.

b) Over-estimated diversity by iTag sequences can result from hypervariability over the V4-region compared to the full-length sequence.

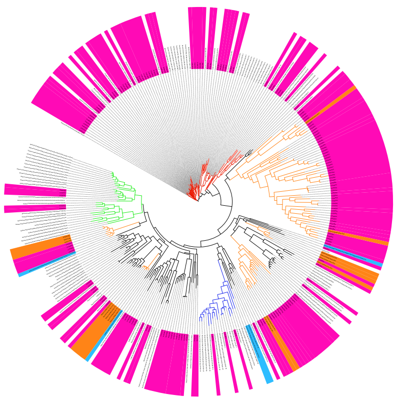

Figure 4: PacBio OTU tree (359 assigned OTUs). Pink: 175 (48.74%) unclassified OTUs; orange text labels: candidate phyla; blue text labels: phyla classified in PacBio data and absent from iTags. Branch colors denote phyla clades.

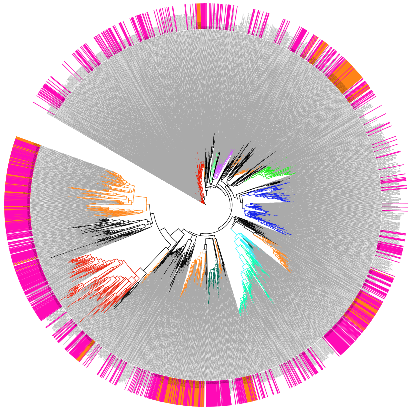

Figure 5: iTag OTU tree (1,843 assigned OTUs). Pink: 828 unclassified OTUs (44.9%); orange text labels: candidate phyla; branch colors denote phyla clades.

Materials & Methods

DNA from species in the mock community was obtained from the American Type Culture Collection (ATCC). The environmental water sample was collected at 120 m depth from Sakinaw Lake, British Columbia, Canada, in 2010. DNA was isolated using 0.22 µM Sterivex Filters and Cesium Chloride Density Gradient Centrifugation [1]. For universal amplification of the V4 region of the 16S rDNA (iTags), forward primer 515F and reverse primer 806R were used (250 bp insert). Full-length 16S rDNA amplification was performed using primers 27F and 1492R. DNA was amplified using the KAPA SYBR FAST qPCR Kit. Pooled amplicons were purified with AMPure (Agencourt Bioscience, Beverly, MA) and analyzed with a Bioanalyzer 2100 (Agilent) instrument. Amplicons and genomic DNA were sequenced on an Illumina Miseq, HiSeq 2500, and a PacBio RSII sequencing platform, respectively. Fro the mock community, we obtained 2,116,448 iTag reads, 7,438,720 Illumina shotgun reads, ~10,000 PacBio 16S reads, and 604,529 PacBio Shotgun reads.

16S sequences of the Sakinaw sample were first rarefied to 5,000 reads. OTUs with ≥ 2 reads were used for phylogenetic analysis resulting in 4,407 iTag and 5,000 PacBio sequences. Sequences were initially aligned using the SINA aligner [2]. Phylogeny based on this alignment was reconstructed using fasttree [3].

Conclusions

1) The two closely related *Salmonella* spp. were not resolved by iTag sequences, because dissimilarities occur outside the V4-region
2) PacBio 16S sequences do not appear to be as influenced by %GC-content as Illumina sequences
3) There are minor differences between datasets from iTags and PacBio 16S in the classifiable OTUs
4) iTag sequences may overestimate microbial diversity as 16S variability in the V4-region is not homogeneous across phyla
5) Full-length 16S sequences are essential for the expansion of current databases because partial 16S can be more accurately placed when few or no cultured representatives are available

References

1: Wright, J. J., Lee, S., Zaikova, E., Walsh, D. A., Hallam, S. J. (2009) DNA Extraction from 0.22 µM Sterivex Filters and Cesium Chloride Density Gradient Centrifugation. J Vis Exp 31: 1352

2: Pruesse, E., Peplies, J. and Glöckner, F.O. (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics, 28, 1823-1829

3: Price, M. N., Dehal, P. S., Arkin, A. P. (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PloS ONE 5 (3)

Acknowledgement/Contact

For questions or comments, please contact:
Esther Singer: esinger@lbl.gov or
Devin Coleman-Derr: dacoleman-derr@lbl.gov

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.