EEB-SATURDAY-1109
17th of June, 2023

Increasing throughput of full-length 16S sequencing using concatenation

Jacob Brandenburg[1], Khi Pin Chua[1], Siyuan Zhang[1], Jason Underwood[1], Minning Chin[1], Wei-Shen Cheng[1], Sian Loong Au[1], Primo Baybayan[1], Holly Ganz[2], Guillaume Jospin[2], Ye Tao[3], Qin Lin[3], Elizabeth Tseng[1], Jeremy E Wilkinson[1]
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025 2. AnimalBiome, 400 29th St., Ste 101, Oakland, CA, USA 94609 3. Biozeron Biotechnology Co., Ltd., Shanghai, 201800 China

PacBio

jwilkinson@pacb.com

Background

In the past several years, the ability to capture full-length 16S rRNA gene with PacBio HiFi sequencing has enabled researchers to profile microbiomes in significantly higher resolution. Only full-length and highly accurate 16S sequences can robustly identify the broad range of bacteria seen in complex microbial communities at the species level, without bias.

To further increase the cost effectiveness of full-length 16S sequencing, we applied the multiplexed array sequencing (MAS-Seq) method (Al'Khafaji et al., 2021, Biorxiv) to 16S amplicons. The MAS-Seq method is a versatile throughput increase method that takes advantage of the longer HiFi read lengths to concatenate smaller amplicons into ordered arrays with programmable array sizes. We demonstrated that MAS-Seq could be applied to 16S amplicons with an ~8 – 12-fold throughput increase. We tested the method on a diverse range (11 types) of samples including mock communities, human and animal feces/guts, soil, and water.

We then analyzed the data using a user-friendly bioinformatics pipeline based on Nextflow called "pb-16S-nf" that was developed to provide a FASTQ-to-report analysis solution for full-length 16S HiFi reads (MAS and non-MAS). Comparing the concatenated (16S MAS-Seq) to non-concatenated full-length 16S datasets, we found no bias in community compositions and were able to assign up to ~90 – 99% of denoised reads to species. In addition, on the highly complex ZymoBIOMICS Fecal Reference with TruMatrix Technology (D6323) sample, we found 16S MAS-Seq to have high correlation to taxonomic abundances estimated from shotgun metagenomics sequencing using the same sample, emphasizing that it's possible to get shotgun metagenome taxonomic resolution at amplicon sequencing costs with full-length 16S HiFi sequencing. Furthermore, with 16S MAS-Seq, researchers may now multiplex more samples to reduce cost/sample or to profile each sample deeper with more reads/sample.

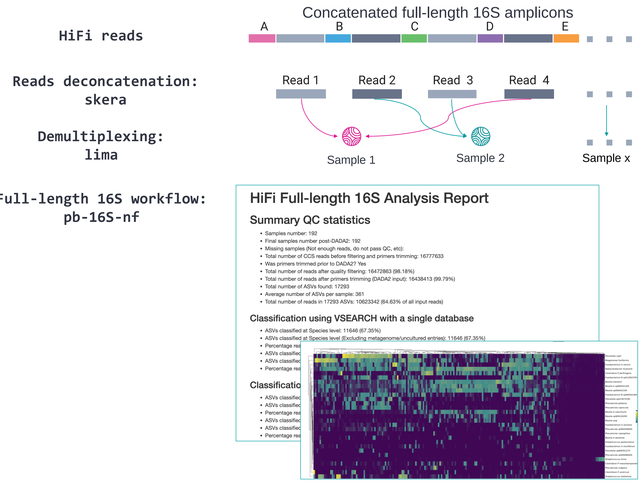

Figure 1. Workflow of 16S MAS-Seq sequencing. Full-length 16S amplicons are concatenated using the MAS-Seq protocol with an array size of 12. The reads are then deconcatenated into individual full-length 16S reads followed by demultiplexing using samples barcodes. Finally, pb-16S-nf is used to process and analyze the reads to generate amplicon sequence variants (ASV) and estimate the relative abundances of different species in the samples.

16S MAS-Seq sequencing provides more HiFi reads

Using the PacBio Sequel IIe system, we partnered with AnimalBiome and Biozeron Biotechnology to generate between 15m to 22m HiFi reads with one SMRT Cell 8M across four sequencing runs. The datasets include sample types such as human and animal feces/guts, food, soil, seawater, and others. Applying pb-16S-nf to these datasets, we observe >80% reads assigned to species level for human and insect guts, fermentation, sediment, cat and dog fecal and oral sample types. All four sequencing runs achieved >15m HiFi reads after deconcatenation and provided significant reads/sample at high (384 samples) multiplexing level.

Experiment	Samples	HiFi reads	Mean QV	S-reads	Mean array size	Reads/sample
Zymo D6323	192	1,874,084	Q33	22,071,938	11.78	**107,533**
AnimalBiome	192	1,617,357	Q35	17,147,299	11.00	**87,383**
Biozeron Run 1	384	1,502,307	Q35	17,235,123	11.47	**44,883**
Biozeron Run 2	384	1,454,210	Q33	15,196,011	10.45	**39,573**

Table 1. Sequencing metrics for 16S MAS-Seq. High quality HiFi sequencing data was generated for all four sequencing runs including many different sample types.

Using samples from AnimalBiome, we compared 16S MAS-Seq sequencing data to non-concatenated 16S data (standard full-length 16S sequencing) and found >97% correlation in Shannon diversity between the two sequencing methods (Fig. 2A). In addition, we compared relative abundance of species estimated based on shotgun-metagenomics sequencing to that of 16S MAS-Seq sequencing and found high correlation between the two methods (Fig. 2B)

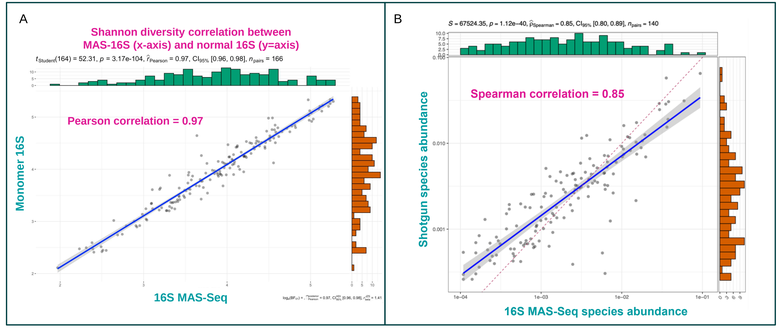

Figure 2. Checking for biases in full-length 16S data. (A) For each sample in the AnimalBiome dataset, Shannon diversity is calculated using data from MAS-Seq and standard full-length 16S sequencing. MAS-16S samples' depths were downsampled to similar depth as monomer 16S. Figure shows that MAS-Seq provides diversity estimate close to standard full-length 16S sequencing. (B) Figure shows high correlation of relative abundance estimates at species level between 16S MAS-Seq and shotgun metagenomics sequencing method using the Zymo D6323 sample.

The pb-16S-nf workflow enables scalable full-length 16S data analysis

We developed an ASV-based analysis workflow written in Nextflow to analyze full-length 16S HiFi sequencing data. The software used in the pipeline are containerized for ease-of-installation and include well-established tools such as Qiime 2 and DADA2. pb-16S-nf utilizes Nextflow to scale analysis of large number of samples on high performance computing cluster and can be adopted to cloud with relatively minor efforts. pb-16S-nf implements a simply VSEARCH-based (GTDB r207) and an exhaustive naïve-Bayes classification method (Multiple databases, see GitHub's FAQ for more details). We ran pb-16S-nf on 16S MAS-Seq data across different sample types and found high percentage or reads assigned at species level. The relatively low amount of reads assigned at species level for complex environmental samples highlights the need for sample-type-specific databases which may improve taxonomy resolution (e.g. MiDAS database for activated sludge increased percentage reads classified at species level from 38% to 89% (Figure 3) using VSEARCH).

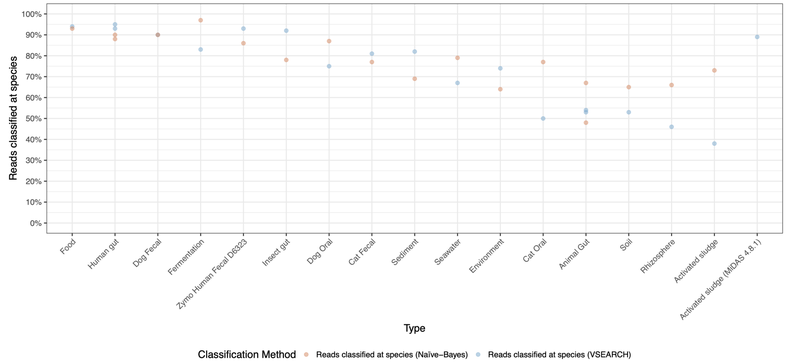

Figure 3. Percentage of reads classified at species level. After running pb-16S-nf, we calculate the percentage of denoised reads that belong to ASVs assigned at species level. For activated sludge, we also ran VSEARCH taxonomy classification using MiDAS 4.8.1.

Summary

- The MAS-Seq method enables significant throughput increase for full-length 16S HiFi sequencing.
- MAS-Seq 16S sequencing does not show bias compared to standard full-length 16S sequencing.
- Relative species abundance estimated based on MAS-Seq 16S sequencing are highly concordant with estimates based on shotgun metagenomics sequencing.
- Up to ~90 – 99% of denoised full-length 16S reads were able to be assigned to species, dependent on the sample type and reference database

References / More information

- pb-16S-nf workflow: https://github.com/PacificBiosciences/pb-16S-nf
- https://www.pacb.com/wp-content/uploads/Application-Brief-Metagenomic-Sequencing-with-HiFi-Reads-Best-Practices.pdf
- https://www.pacb.com/wp-content/uploads/metagenomics-solutions-brochure.pdf
- https://www.pacb.com/products-and-services/applications/complex-populations/microbial/