High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution

Benjamin J Callahan[1,2,*], Joan Wong[3], Cheryl Heiner[3], Steve Oh[3], Casey M Theriot[1], Ajay S Gulati[4,5,6], Sarah K McGill[7], Michael K Dougherty[7]

1. Department of Population Health & Pathobiology, North Carolina State University, Raleigh, NC 27607
2. Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695
3. Pacific Biosciences of California, Inc., Menlo Park, CA 94025
4. Center for Gastrointestinal Biology and Disease, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
5. Department of Pediatrics, Division of Gastroenterology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
6. Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
7. Department of Medicine, Division of Gastroenterology and Hepatology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

*: Correspondence should be addressed to benjamin.j.callahan@gmail.com

Author Contributions

BJC designed the research; BJC implemented the algorithm; BJC performed the analysis; BJC wrote the paper; JW, CH and SO developed the amplicon sequencing methodology, performed the amplicon sequencing, and processed the raw sequencing data; CMT, ASG, SKM and MKD collected the human fecal samples.

Conflict of Interest Statement

The sequencing data investigated in this manuscript were generated by Pacific Biosciences Inc, Menlo Park, CA. JW, CH and SO are full-time employees at Pacific Biosciences, a company commercializing single-molecule sequencing technologies.

Abstract

Targeted PCR amplification and high-throughput sequencing (amplicon sequencing) of 16S rRNA gene fragments is widely used to profile microbial communities. New sequencing technologies produce long reads that can span the entire 16S rRNA gene, but have substantially higher error rates that have limited their attractiveness when accuracy is important. Here we present a high-throughput amplicon sequencing methodology based on PacBio circular consensus sequencing and the DADA2 sample inference method that measures the full-length 16S rRNA gene with single-nucleotide resolution and a near-zero error rate.

In two artificial mixtures of known bacterial strains our method recovered the full complement of full-length 16S sequence variants from expected community members, without residual errors. The measured abundances of intra-genomic sequence variants were in the integral ratios expected from the genuine allelic variants within a genome. *E. coli* strains in the mock communities were correctly classified to the O157:H7 and K12 sub-species clades from the 16S gene sequences recovered by our method. In human fecal samples, our method recovered the full complement of 16S rRNA gene variants in detected *E. coli* strains and showed strong technical replication.

We discuss the promises and challenges of of classification based on the full complement of multi-copy marker genes such as the 16S rRNA gene. There are likely many applications beyond microbial profiling for which high-throughput amplicon sequencing of complete genes with single-nucleotide resolution will be of use.

Introduction

The amplification of specific genetic loci by polymerase chain-reaction (PCR) can powerfully focus DNA sequencing on genetic variation of interest. Amplicon sequencing effectively detects genetic variation embedded in complex chemical and genetic backgrounds, and is far more cost-effective than untargeted sequencing when large amounts of undesired genetic material is present, as can be the case for host-associated microbial populations or specific genes in large genomes (Franzosa 2015). The precision, sensitivity and low cost of amplicon sequencing have made it a ubiquitous tool utilized in thousands of published scientific studies each year.

However, the advantages of amplicon sequencing come at the cost of a sharply limited genetic field of view. In current practice, the genetic loci measured by amplicon sequencing are typically restricted to 100–500 nucleotide regions that fit within the short reads generated by high-throughput sequencing platforms. In the popular community profiling application, short reads limit investigators to fragments of preferred taxonomic barcodes, such as the 16S rRNA gene in bacteria or the ITS region in fungi, degrading taxonomic resolution and the ability to distinguish between related strains (Fuks 2018; Edgar 2018). In studies of functional genes, short reads do not cover even compact viral genes, limiting amplicon sequencing to measurements of incomplete gene slices.

In recent years, Pacific Biosciences (PacBio) and Oxford Nanopore have developed new technologies that generate long sequencing reads that can extend tens of thousands of nucleotides (Goodwin 2016; Levy 2016). Long reads can dramatically widen the genetic field of view measured by amplicon sequencing, offering the promise of greatly increased resolution in taxonomic profiling applications and measurement of complete functional genes. However, amplicon sequencing applications are often sensitive to the presence of spurious sequence variants introduced by PCR and sequencing errors, and long-read sequencing has a much higher error rate (~10%) than does short-read sequencing (~0.5%). For PacBio, high long-read error rates can be ameliorated by the construction of circular consensus sequences (CCS), in which an amplified genetic locus is circularized and read through multiple times before a consensus sequence is reported (Hebert 2018). The CCS approach effectively trades read length for accuracy: CCS read lengths extend only 1-10 kilobases, but per-base accuracy has been reported to be comparable to short-read sequencing (Jiao 2013; Larsen 2014).

Several previous studies have evaluated long-read amplicon sequencing of the 16S rRNA gene based on PacBio CCS and Oxford Nanopore reads, but have reported that a still considerable error rate necessitated lumping similar sequences together to mitigate the impact of sequencing errors on subsequent analyses (Schloss 2016; Singer 2016; Wagner 2016; Schlaeppi 2016; Calus 2018). As a result, long-read amplicon sequencing was not found to improve resolution relative to short reads as much as was suggested by the increase in read length. The advantages of long reads are a sufficient value proposition in certain applications, such as tracking diversification of the HIV *env* gene (Caskey 2017; Eren 2017) and characterizing

histocompatibility in non-human organisms (Westbrook 2015; Karl 2017), but in practice the limited resolution gain has been insufficient to justify the increased cost of long-read amplicon sequencing in many applications, and so it is not widely used at this time.

Here we introduce an amplicon sequencing methodology based on PacBio CCS sequencing and the DADA2 algorithm and software (Callahan 2016) that captures the full-length 16S rRNA gene with single-nucleotide resolution and a near-zero error rate. We demonstrate the accuracy and precision of our method on mock microbial communities and on technically replicated samples from the human fecal microbiome. Although our focus here is on the 16S rRNA gene, the core components of this methodology are extensible to other genetic loci for which suitable primer sets are available.

Results

Mock Communities

We evaluated the performance of our full-length 16S rRNA amplicon sequencing methodology in two artificially constructed communities of known composition (mock communities). The Zymo mock community contains eight phylogenetically distant bacterial strains, with cellular proportions chosen to equalize the total genomic DNA contributed by each strain. The HMP mock community contains 20 bacterial strains of varying phylogenetic similarity, with proportions chosen such that strain rRNA gene frequencies vary over 3 orders of magnitude. Each mock community was sequenced on a single Sequel cell. They Zymo mock community yielded 77,453 CCS reads above the *minPredictedAccuracy* threshold of 99.9%, and 69,367 reads after removing primers and filtering (Methods). The HMP mock community yielded 78,328 reads above the threshold, and 69,963 reads after removing primers and filtering.

Exact amplicon sequence variants (ASVs) were inferred from the filtered reads by a new version of the DADA2 method updated to efficiently process long amplicon reads and appropriately model PacBio CCS sequencing errors (Methods). Taxonomy was assigned to ASVs by the naive Bayesian classifier and the SILVA v128 database (Wang 2007; Quast 2013). Species-level assignments were made by BLAST searches against the NCBI nucleotide collection (nt) for the multiple species of *Staphylococcus* and *Streptococcus* in the HMP mock community that were otherwise identical at the genus level. In both datasets, every ASV was assigned to a genus or species belonging to the expected members of the mock communities. Since we expect to detect intragenomic variation in the often multi-copy 16S rRNA gene, we grouped all ASVs into putative strain bins by their genus or species assignments.

In the Zymo mock community 29 ASVs were reported, of which 25 of 29 were exact matches (100% identity, 100% coverage) to 16S rRNA genes previously sequenced from the isolates of the corresponding species. The three *Lactobacillus fermentum* ASVs and one *Staphylococcus aureus* ASV that were not exact matches differed by just one nucleotide from previously

observed sequences. In the HMP mock community 51 ASVs were reported, of which 48 of 51 were exact matches. The one *Staphylococcus epidermis* ASV, and two *Clostridium bejerenckii* ASVs that were not exact matches differed by just two and one nucleotides from previously observed sequences, respectively. The frequencies of ASVs detected in the HMP mock community vary over 3 orders of magnitude, and range as low as 0.00019.

If ASVs represent genuine allelic variants present in these genomes, then ASVs assigned to the same strain should appear in integral ratios to one another, whereas integral ratios are unexpected if ASVs represent uncorrected amplicon sequencing artefacts. The ASVs inferred by our methodology appear in frequencies that are integral ratios to one another, and to the genomic frequency of the corresponding strain. Figure 1 shows the frequency of each ASV in the Zymo mock community, normalized to the observed strain genomic frequency (Methods). Every ASV:genome ratio is an integral value (1, 2, 3, …), with a maximum deviation of ±0.2. The frequencies of ASVs in the HMP mock community show the same pattern of clear integral ratios, with the exception of a deviation in *Bacillus cereus* ASV frequencies that cannot be attributable to counting noise, given the substantial frequency at which that strain is present in the data (Figure 2).

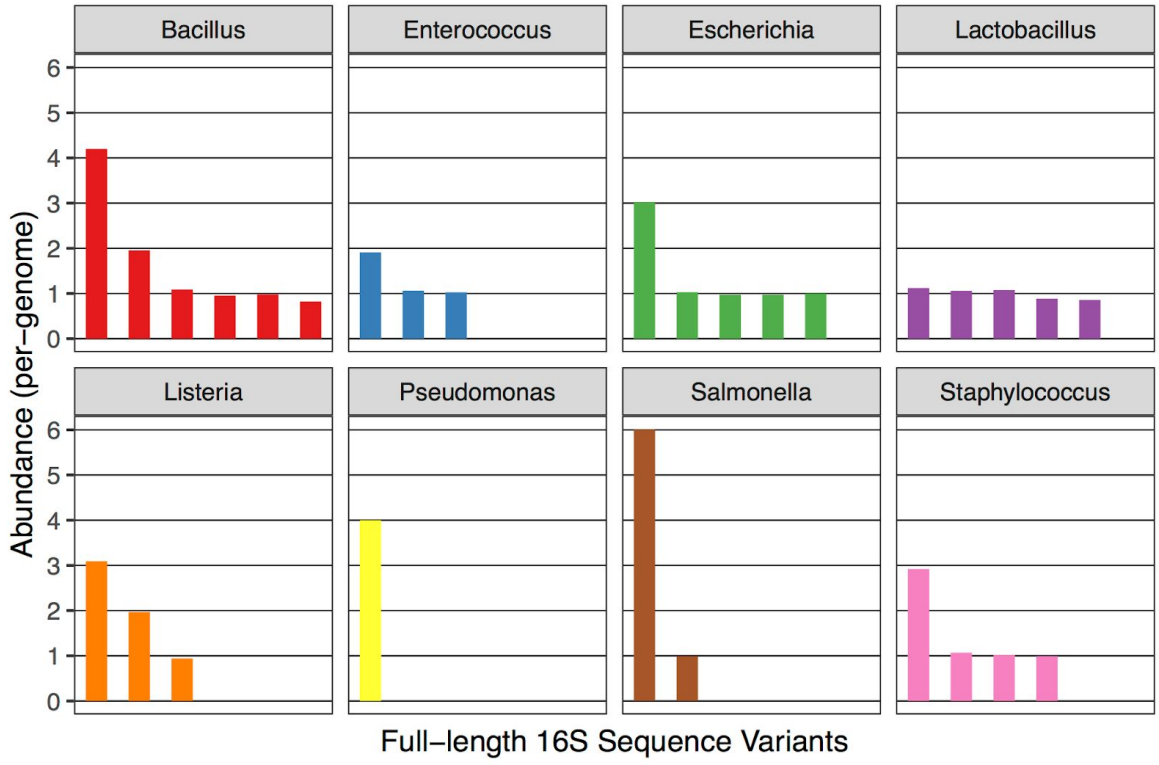

Figure 1. Frequencies of ASVs recovered from the Zymo mock community, scaled by the genomic frequency. The abundance of each ASV divided by the genomic frequency of the mock community strain from which it originated is plotted on the y-axis. Integral values are indicated by horizontal grid lines. Each facet corresponds to one of the 8 bacterial strains in the Zymo mock community. No other ASVs were reported.

Earlier evaluations of full-length 16S rRNA gene amplicon sequencing using PacBio CCS reads generated by the RSII sequencing instrument reported the presence of systematic errors, i.e. errors that repeatedly arose at specific positions (Schloss 2016; Wagner 2016). The existence of systematic errors would be inconsistent with the zero or near-zero error rates of the ASVs in these mock communities, as the DADA2 method would interpret sufficiently repeated errors as biological variation. These earlier reports may have reflected unrecognized intragenomic variation. We re-analyzed the sequencing data from the *Staphylococcus aureus* monoculture data used to describe systematic errors in Wagner et al. 2016 with DADA2. We recovered 5 ASVs, all of which exactly matched previously sequenced 16S rRNA genes from *S. aureus*, whose genome typically contains 5 copies of the *rrn* operon. The differences between these intragenomic variants may have been misinterpreted as systematic errors, perhaps because the short-read genome assembly that was used as the ground truth contained only one of the five *rrn* operons in the *S. aureus* genome (Larner-Svensson 2013).

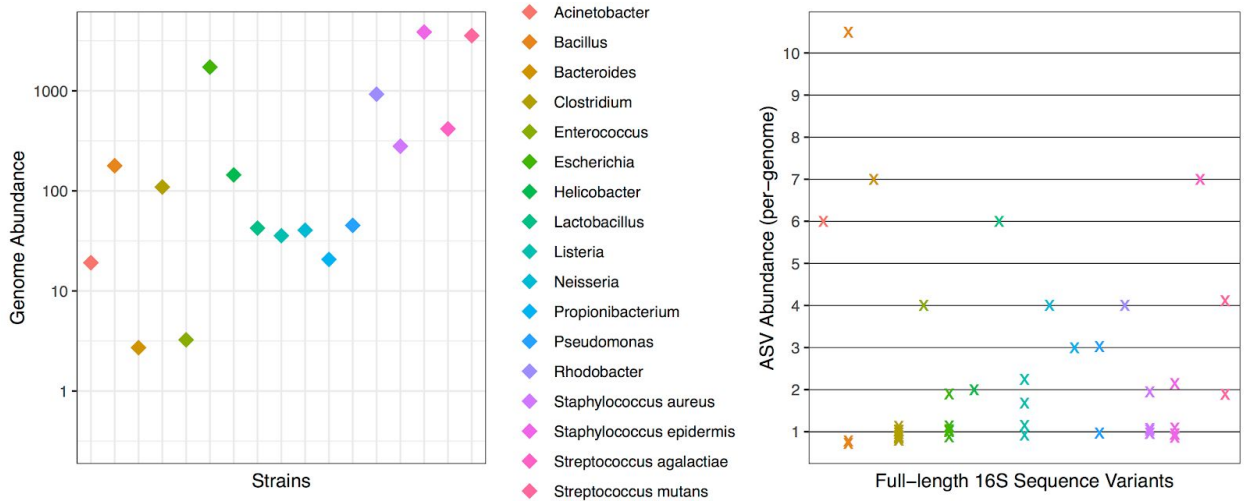

Figure 2. Frequencies of genomes and ASVs recovered from the HMP mock community. (a) The abundances of each genome in the data are plotted on the log-scaled y-axis (Methods). There is variation in genome frequency over three orders of magnitude, and significant counting noise exists for ASVs from genomes with abundances below 100. (b) The abundance of each ASV divided by the genomic frequency of the mock community strain from which it originated is plotted on the y-axis. Integral values are indicated by horizontal grid lines. No other ASVs were reported.

Sub-species Classification by the Full Complement of 16S rRNA Gene Alleles

Typically taxonomic classification is performed on individual 16S rRNA gene sequences, but even greater resolution can be obtained by using a strain's full complement of 16S alleles. We demonstrate the resolution achievable by such an approach on the *E. coli* strains in each mock community. In the Zymo mock community, 5 *E. coli* ASVs were recovered with abundances in 3:1:1:1:1 ratios. In the HMP mock community, 6 *E. coli* ASVs were recovered with abundances in 2:1:1:1:1:1 ratios. We consider here a simple *ad hoc* procedure, in which we collect

accessions for the best BLAST hits of individual ASVs, and then intersect the best-hit accessions recorded for each ASV to find the set of maximally matching accessions (Methods).

The *E. coli* strain in the Zymo mock community could be classified as belonging to the O157:H7 clade based on its full complement of 16S rRNA genes. There were 14 accessions that exactly matched 4 of the 5 ASVs recovered from the Zymo *E. coli* strain, and none that exactly matched all five. Of those 14 accessions, 12 were annotated with serotype O157:H7, one was annotated with O antigen 157 but had no annotation for the H antigen, and one had no serotype annotation. Leaving aside the missing values, the Zymo *E. coli* strain was unambiguously and correctly classified as a member of the O157:H7 clade.

The *E. coli* strain in the HMP mock community could be classified as belonging to the K12 clade based on its full complement of 16S rRNA genes. There were 32 accessions that exactly matched all 6 ASVs recovered from the HMP *E. coli* strain. Of those 32 accessions, 27 were either annotated as K-12, MG1655, or as derived from MG1655 (a specific strain in the K-12 clade). The other 5 are unannotated, but the best BLAST hits to the full genomes are K-12 strains. Leaving aside the missing values, the HMP *E. coli* strain was unambiguously and correctly classified as a member of the K-12 clade.

Human Fecal Samples

We next sought to validate the performance of our amplicon sequencing methodology in complex microbial communities. To do so, we performed replicate characterizations of a set of 12 human fecal samples. The full-length 16S rRNA gene was amplified, each sample was barcoded, and multiplexed sequencing was performed in a single Sequel cell. The first replicate was sequenced with the S/P2-C2/5.0 Sequel sequencing chemistry, which yielded 177,691 CCS reads at 99.9% predicted accuracy across the twelve samples and 146,589 reads after primer detection and filtering, for a median of 12,461 filtered reads per sample. The second replicate was sequenced with a pre-release version of S/P3-C3/5.0 Sequel sequencing chemistry, which yielded 289,644 CCS reads at 99.9% predicted accuracy across the twelve samples and 249,802 reads after primer detection and filtering for a median of 20,284 filtered reads per sample, nearly double that achieved by the currently available sequencing chemistry.

We do not know the ground truth of these natural community samples, so as an alternative to accuracy analysis we investigated consistency across technical replicates. Prior to sample inference with DADA2, we rarefied each replicate sample to 10,000 reads in order to remove any impact of sequencing depth. The ASVs detected by DADA2 from these rarefied samples were consistent across technical replicates (Figure 3A). On a per-sample basis, 1086 ASVs were detected in both replicates, while 115 and 128 were detected only in Replicate 1 or 2, respectively. Estimated abundances were highly consistent: the Pearson's correlation between the per-sample abundances across replicates was 0.998. ASVs that failed to replicate appeared at low frequencies (<50 reads, <0.5%; Figure 3A).

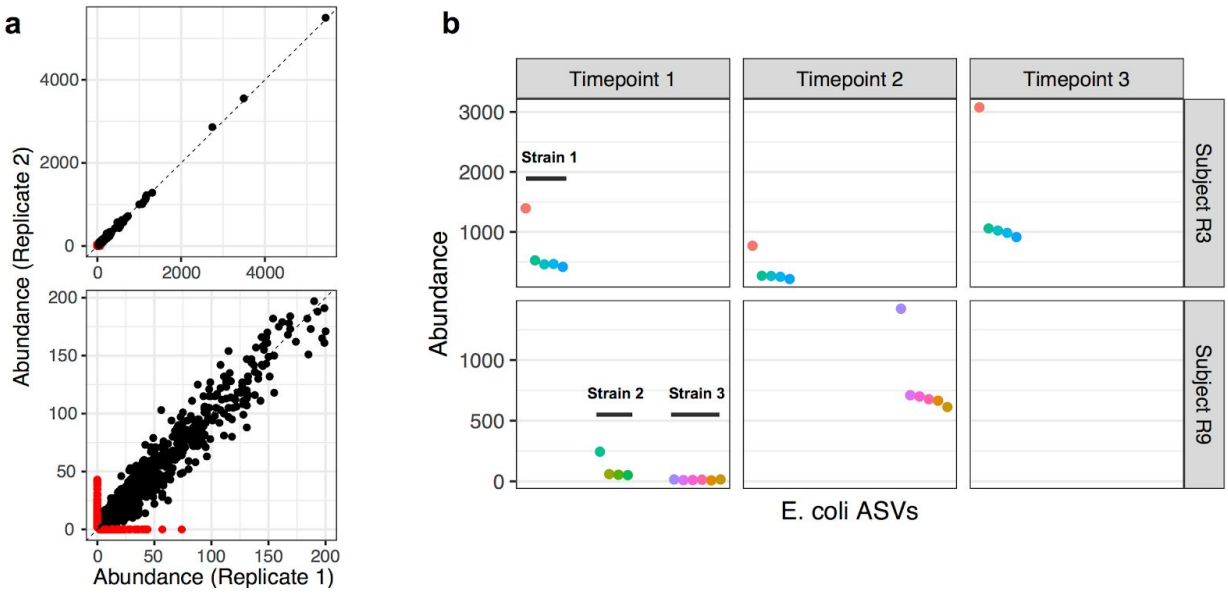

Figure 3. Consistent detection of ASVs in human fecal samples. 12 human fecal samples were characterized by two technical of our full-length 16S rRNA gene amplicon sequencing method. (a) The abundances of all ASVs recovered from the same sample in different technical replicates. ASVs recovered in both technical replicates are black, and non-replicated ASVs are red. The top panel shows the full range of per-sample abundances, and the bottom panel zooms in on low abundance ASVs. (b) The abundance of each ASV assigned to *E. coli* in Replicate 2 for samples with >0.2% *E. coli* reads. Longitudinal samples from subjects R3 (top row) and R9 (bottom row) are ordered left-to-right by sampling time. ASVs sharing a putative strain-level bin are plotted adjacently.

To investigate whether the full complement of 16S rRNA gene sequence variants within individual strains could be resolved in more complex natural communities, we focused on all ASVs in the fecal samples that were classified as *Escherichia coli*. In each sample in which an appreciable number of *E. coli* reads was detected, clear strain-level bins could be constructed based on the expected integral ratios between the abundances of intra-genomic alleles, as well as our knowledge that *E. coli* has 7 copies of the 16S rRNA gene (Figure 3B). Consistent strain-level bins persist within subjects over time. In the Timepoint 1 sample from Subject R9, two different *E. coli* strains can be clearly distinguished from one another. The full complement of ASVs from the low-frequency *E. coli* strain in this sample was recovered, even though individual ASVs were present in only 8–15 reads each. Furthemore, these low frequency ASVs exactly match the ASVs from the strain present in the Timepoint 2 sample.

Discussion

Currently, community profiling of the 16S rRNA gene is conducted using short-read sequencing technologies that measure only fragments of the gene, substantially degrading accuracy and resolution. This is largely responsible for the well-known difficulty in achieving species-level resolution from high-throughput 16S sequencing data (Edgar 2018). The potential benefits of

capturing the entire 16S rRNA gene are clear (e.g. Earl 2018), but higher costs and error rates as well as limited computational methods have limited the appeal of long-read amplicon sequencing to date. We believe that recent and ongoing improvements in PacBio sequencing instruments and chemistries, coupled with the accurate and high-resolution computational methods we are presenting here, make a compelling case for investigators to revisit long-read amplicon sequencing as a measurement technique going forward.

One of the challenges in evaluating our method was that the accuracy and completeness of the ASVs we recovered often outstripped supposedly authoritative references. This was particularly true when reference genome assemblies were created by short-read sequencing, which struggles to resolve repeated regions such as multiple copies of the *rrn* operon. As a result, we used a multifaceted approach in which we compared ASVs to the references provided with the mock community materials and to broader reference databases such as nt, and we also investigated the pattern of integral ratios that is expected between genuine allelic variants within a genome. In the Zymo mock community, we conclude from this approach that output of our method contained no false positives and had a zero residual per-base error rate. In the HMP mock community, we conclude that a residual per-base error rate of zero is most likely, but we leave open the possibility that the *B. cereus* variants that did not match expected integral ratios could be errors, in which case the residual per-base error rate would be 2.6×10^{-6}. While we cannot directly measure accuracy in the human fecal samples, strong technical replication and clear recovery of the full complement of 16S rRNA genes from multiple *E. coli* strains are consistent with high accuracy over a wide frequency range in more complex samples. These results suggest that this methodology may be useful for generating reference-quality gene sequences, particularly for multi-copy genes. The short turn-around time of PacBio sequencing and the effectiveness of amplicon sequencing in mixed and heterogeneous samples also suggest potential diagnostic applications (Cummings 2016).

The high resolution and accuracy we are reporting derives in part from the exceptional and not-entirely-appreciated accuracy of PacBio CCS sequencing. In the mock community datasets, we find that half of all sequencing reads are error-free over their entire ~1.5 kilobase (kb) length. As a result, a computational approach leveraging repeated observations of the error-free sequence was adaptable to PacBio CCS data (Callahan 2016; Callahan 2017). We predict that our DADA2-based computational workflow will continue to be effective for PacBio CCS amplicons extending out to ~3 kb, but will degrade for >3 kb amplicons, especially for lower-frequency variants, as the fraction of error-free reads declines. We also urge caution in applying our computational methods to data from PacBio sequencing chemistries before P6-C4 and/or CCS data that was generated by early versions of the earlier SMRT Portal software, as error rates in such data may be substantially higher than in the data considered here.

Single-nucleotide resolution of the full-length 16S rRNA gene often reveals significant intragenomic heterogeneity. At first glance this may appear to be an unwanted nuisance, at least for simple applications such as counting numbers of microbial objects in a community. However, there is also tremendous opportunity presented by this increased resolution and

accuracy, particularly if new methods can be developed that automate the simple rules we used to create strain-level bins in our data by hand: integral ratios, similar taxonomic assignments, and consistent patterns across samples. There may be opportunities to translate techniques from current metagenomics binning approaches for that purpose (e.g. Alneberg 2014). It also may be possible to augment existing taxonomic classification methods to utilize the full complement of 16S rRNA genes from strain-resolved bins.

Full-length 16S rRNA gene profiling is compelling, but it cannot achieve the same level of resolution that is possible from deep shotgun sequencing, at least for strains abundant enough to be assembled, and it does not speak directly to functional potential. Furthermore, if only coarse taxonomic profiles are desired (e.g. Family level or above), then the lower costs of short-read amplicon sequencing will likely make it the preferred approach. Nevertheless, high-throughput amplicon sequencing of the full-length 16S rRNA gene is an attractive option for a wide variety of applications that benefit from the higher resolution and classification accuracy provided by the full-length gene, from the advantages of targeted amplicon sequencing, e.g. environments in which the genetic material of interest is a minority of all DNA, or from the superior 16S rRNA gene reference databases available, e.g. environments less well characterized than the human gut.

Finally, while we focused here on the 16S rRNA gene, it is not obvious that this is the most compelling application of this technology, and our methodology can be straightforwardly adapted to other genetic loci. Other barcoding genes, such as the ITS region in fungi (Tedersoo 2018), would also clearly benefit from accurate long-read amplicon sequencing with single-nucleotide resolution, and population-level structure can be probed by targeting faster-evolving markers. However, the most intriguing applications may not be barcoding genes at all, but functional genes in which fine-scale variation across the entire gene is important in and of itself. Examples that are already receiving research attention include the *env* gene in HIV-1 and the MHC locus in humans and non-human models (Caskey 2017; Karl 2017), but further possibilities abound.

Methods

Mock Communities

The “Zymo” mock community is the ZymoBIOMICS™ Microbial Community DNA Standard available from Zymo Research (Irvine, CA). The Zymo mock community consists of 8 well-separated bacterial strains, 3 of which are gram-negative and 5 of which are gram-positive, as well as 2 yeast strains not amplified by our amplicon sequencing protocol. Genomic DNA from each bacterial strain is mixed in equimolar proportions.

Of note, Zymo Research replaced five strains in the ZymoBIOMICS™ standards with similar strains beginning with Lot ZRC190633. The Lot # of the sample analyzed here was

ZRC187325, which contains the old mixture of strains, including *E. coli* O157:H7 str. CDC B6914-MS1. Detailed information is available at the manufacturer's website: https://www.zymoresearch.com/zymobiomics-community-standard.

The "HMP" mock community was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Staggered, Low Concentration), v5.2L, for 16S rRNA Gene Sequencing, HM-783D. The HMP mock community consists of 20 bacterial strains, most of which are well-separated from each other, but also contains two *Staphylococcus* species and three *Streptococcus* species. Genomic DNA from each bacterial strain is mixed so that the concentration of the rRNA gene operon varies over three orders of magnitude. Detailed information is available at the BEI website: https://www.beiresources.org/Catalog/otherProducts/HM-783D.aspx.

Fecal Samples

Genomics DNA was extracted from homogenized fecal samples with the MO Bio PowerFecal kit (Qiagen) automated for high throughput on QiaCube (Qiagen). The manufacturer's instructions were followed with bead beating in 0.1 mm glass bead plates. DNA quantification was performed with the Qiant-iT Picogreen dsDNA Assay (Invitrogen).

Amplicon Sequencing

Amplicon sequencing was performed by Pacific Biosciences Inc. (Menlo Park, CA) from genomic DNA extracted from heterogeneous microbial communities. Briefly, the 27F:AGRGTTYGATYMTGGCTCAG and 1492R:RGYTACCTTGTTACGACTT primer set was used to amplify the full-length 16S rRNA gene from genomic DNA. Amplified DNA from each mock community was sequenced on a dedicated PacBio Sequel cell using with the S/P1-C1.2 sequencing chemistry. DNA from the 12 fecal samples was amplified by barcoded 16S primers in order to reduce chimera formation as compared to a two-round PCR approach. Mutiplexed sequencing of the amplified fecal DNA was performed on a single PacBio Sequel cell. Replicate 1 of the fecal samples was sequenced using the S/P2-C2/5.0 sequencing chemistry, and Replicate 2 of the fecal samples was sequenced with a pre-release version of the S/P3-C3/5.0 sequencing chemistry.

Full details are presented in the Full-Length 16S Amplification SMRTbell® Library Preparation and Sequencing Procedure & Checklist available online at https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Full-Length-16S-Amplification-SMRTbell-Library-Preparation-and-Sequencing.pdf.

Bioinformatics

CCS reads in the Zymo and BEI mock communities were generated using the *ccs* application with *minPredictedAccuracy=0.999* in the SMRT Link 3.1.1 software package (Pacific Biosciences, 2016). For the fecal samples, raw reads were first demultiplexed using the lima application, specifying that the same barcodes were attached at both ends of an insert using the flags same and peek-guess, followed by generation of CCS reads using the ccs application with

minPredictedAccuracy=0.999 in the SMRT Link 5.1 software package (Pacific Biosciences, 2018). Of note, the *ccs* application was updated in SMRTLink 3.0 to use the superior Arrow model (Hepler 2016), and CCS reads generated by the earlier SMRT Portal software may not be as accurate.

The DADA2 method was originally developed for short-read amplicon sequencing, and a detailed description of the algorithm is available in the original publications (Rosen 2012; Callahan 2016). The DADA2 software implements a complete workflow that takes raw amplicon sequencing data in fastq files as input, and produces an error-corrected table of the abundances of amplicon sequence variants in each sample (an ASV table) as output. The standard processing steps in the DADA2 workflow include quality filtering, dereplication, sample inference, chimera removal, and taxonomic assignment.

Several improvements were made to enable precise ASV inference from long amplicon reads by version 1.9.1 of the DADA2 software used here. First, the core data structures and sequence comparison algorithms used by the DADA2 algorithm were augmented to handle quality scores up to 93, sequence lengths up to 3000 nucleotides, and variable sequence lengths. A new error estimation procedure was introduced for PacBio CCS data, in which estimation of the error rate for bases with the maximum quality score of 93 is separated from estimation of error rates over the remainder of the quality score distribution. A new *removePrimers* command was added to the dada2 R package, which removes primers from PacBio CCS reads and can also orient the reads in the forward direction, as PacBio CCS reads are generated in a mixture of forward and reverse-complement orientations. A new default value of *minFoldParentOverabundance=4.*5 is recommended for full-length 16S rRNA gene sequencing, in order to avoid spurious identification of some 16S variants as chimeras of other 16S variants that occur at higher copy number within the same genome.

Reproducible R markdown documents implementing the analyses performed in this paper also document the full DADA2 workflow used to infer ASVs from PacBio CCS data: https://github.com/benjjneb/LRASManuscript. All computation was performed on a 2017 MacBook Pro, and was completed in minutes to tens of minutes.

Full-Complement Classification of *E. coli*

ASVs assigned to the *Escherichia* genus were grouped into strain-level bins based on the expectation of integral ratios between same-genome ASVs, as well as the known copy number of seven 16S rRNA genes in *E. coli*. We performed BLAST searches for each ASV in a strain-level bin against nt on August 7, 2018. The highest BLAST score for each ASV was determined, and all accessions reaching the highest score were recorded. The total occurrences of accessions across the high scores for each ASV were tabulated. The set of accessions with the greatest number of high-score hits served as the basis for further classification. The metadata associated with the Genbank entry for each such accession was inspected, and *E. coli* strains were further assigned metadata values (e.g. serotype) that were

consistent across all such accessions. Accessions for which the metadata entry of interest was absent were ignored.

Determination of Genomic Abundances

The genomic rRNA gene copy number of each strain in the Zymo mock community was provided by the vendor. The 16S rRNA gene copy number of each strain in the HMP mock community was determined by reference to the entries corresponding to that species in *rrn*DB version 5.4 (The Ribosomal RNA Database, https://rrndb.umms.med.umich.edu/, Stoddard 2015). Ambiguities were resolved by analyzing the number of 16S rRNA gene copies in the top BLAST hits of the ASVs assigned to each strain. We then used this copy number, along with the frequencies of the associated ASVs, to infer the genomic frequency of each mock community strain in the pool of amplified DNA. More precisely, to calculate the genomic frequencies we summed the frequencies of all ASVs assigned to each strain, and then divided by the corresponding 16S rRNA gene copy number.

Acknowledgements

We thank Naga Betrapally for assistance in generating CCS reads, Meredith Ashby for advice and consultation, and Paul Hess, Roey Angel and Jacob Price for sharing test datasets not included in this manuscript. CoreBiome, Inc. performed the fecal sample DNA extraction and quantification.

Data Availability

Reproducible R markdown documents implementing the analyses presented in this manuscript are available at https://github.com/benjjneb/LRASManuscript. Sequencing data is deposited at the SRA under accession: PENDING

References

Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. (2014). Binning metagenomic contigs by coverage and composition. Nature methods, 11(11), 1144.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods, 13(7), 581-583.

Callahan BJ, McMurdie PJ, Holmes SP. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal, 11(12), 2639.

Calus ST, Ijaz UZ, Pinto AJ. (2018). NanoAmpli-Seq: A workflow for amplicon sequencing from mixed microbial communities on the nanopore sequencing platform. bioRxiv, 244517.

Caskey M, Schoofs T, Gruell H, Settler A, Karagounis T, Kreider EF, Murrell B, Pfeifer N, Nogueira L, Oliveira TY, Learn GH. (2017). Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. Nature medicine, 23(2), 185.

Cummings LA, Kurosawa K, Hoogestraat DR, SenGupta DJ, Candra F, Doyle M, Thielges S, Land TA, Rosenthal CA, Hoffman NG, Salipante SJ. (2016). Clinical next generation sequencing outperforms standard microbiological culture for characterizing polymicrobial samples. Clinical chemistry, 62(11), 1465-73.

Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, Palmer JN, Workman AD, Blasetti M, Hammond J, Cohen NA. (2018). Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. bioRxiv, 338731.

Edgar RC. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics, 1, 5.

Eren K, Weaver S, Ketteringham R, Valentyn M, Smith ML, Kumar V, Mohan S, Pond SL, Murrell B. (2017). Full-Length Envelope Analyzer (FLEA): A tool for longitudinal analysis of viral amplicons. bioRxiv, 230474.

Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nature Reviews Microbiology, 13(6), 360.

Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome, 6(1), 17.

Goodwin S, McPherson JD, McCombie WR. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333.

Hebert PD, Braukmann TW, Prosser SW, Ratnasingham S, Ivanova NV, Janzen DH, Hallwachs W, Naik S, Sones JE, Zakharov EV. (2018). A Sequel to Sanger: amplicon sequencing that scales. BMC genomics, 19(1), 219.

Hepler NL, Delaney N, Brown M, Smith ML, Katzenstein D, Paxinos EE, Alexander D. (2016). An Improved Circular Consensus Algorithm with an Application to Detect HIV-1 Drug-Resistance Associated Mutations (DRAMs). Poster presentation.

https://www.pacb.com/wp-content/uploads/improved-circular-consensus-algorithm-with-application-to-detection-hiv-1-drams.pdf

Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, Sherman BT, Hu X, Jones K, Raley C, Tran B. (2013). A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. Journal of data mining in genomics & proteomics, 4(3).

Karl JA, Graham ME, Wiseman RW, Heimbruch KE, Gieger SM, Doxiadis GG, Bontrop RE, O'Connor DH. (2017). Major histocompatibility complex haplotyping and long-amplicon allele discovery in cynomolgus macaques from Chinese breeding facilities. Immunogenetics, 69(4), 211-229.

Larner-Svensson, H., Worning, P., Bartels, M. D., Hansen, L. H., Boye, K., & Westh, H. (2013). Complete genome sequence of Staphylococcus aureus strain M1, a unique t024-ST8-IVa Danish methicillin-resistant S. aureus clone. Genome announcements, 1(3), e00336-13.

Larsen, P. A., Heilman, A. M., & Yoder, A. D. (2014). The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. BMC genomics, 15(1), 720.

Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. Annual review of genomics and human genetics, 17, 95-115.

Pacific Biosciences. (2018). SMRT® Tools Reference Guide. https://www.pacb.com/wp-content/uploads/SMRT_Tools_Reference_Guide_v510.pdf.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research, 41(D1), D590-6.

Rosen MJ, Callahan BJ, Fisher DS, Holmes SP. (2012). Denoising PCR-amplified metagenome data. BMC bioinformatics, 13(1), 283.

Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ, 4, e1869.

Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, Gies EA, Cheng JF, Copeland A, Klenk HP, Hallam SJ. (2016). High-resolution phylogenetic microbial community profiling. The ISME journal, 10(8), 2020.

Stoddard SF, Smith BJ, Hein R, Roller BR, Schmidt TM. (2014). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic acids research, 43(D1), D593-8.

Tedersoo L, Tooming-Klunderud A, Anslan S. (2018). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist, 217(3), 1370-1385.

Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. (2016). Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC microbiology, 16(1), 274.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16), 5261-7.

Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, Garcia K, Sanchez-Lockhart M, O'Connor DH, Palacios G. (2015). No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. Human immunology, 76(12), 891-896.