A peer-reviewed version of this preprint was published in PeerJ on 28 March 2016.

View the peer-reviewed version (peerj.com/articles/1869), which is the preferred citable publication unless you specifically need to cite this preprint.

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
https://doi.org/10.7717/peerj.1869

Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system

Patrick D Schloss, Sarah L Westcott, Matthew L Jenior, Sarah K Highlander

Over the past 10 years, microbial ecologists have largely abandoned sequencing 16S rRNA genes by the Sanger sequencing method and have instead adopted highly parallelized sequencing platforms. These new platforms, such as 454 and Illumina's MiSeq, have allowed researchers to obtain millions of high quality, but short sequences. These platforms have allowed researchers to significantly improve the design of their experiments. The tradeoff has been the decline in the number of full-length reference sequences that are deposited into databases. To overcome this problem, we tested the ability of the PacBio Single Molecule, Real-Time (SMRT) DNA sequencing platform to generate sequence reads from the 16S rRNA gene. We generated sequencing data from the V4, V3-V5, V1-V3, V1-V6, and V1-V9 variable regions from within the 16S rRNA gene from a synthetic mock community and natural samples collected from human feces, mouse feces, and soil. The synthetic mock community allowed us to assess the actual sequencing error rate and how that error rate changed when different curation methods were applied. We developed a simple method based on sequence characteristics and quality scores to reduce the observed error rate for the V1-V9 region from 2.16% to 0.32%. Unfortunately, this error rate was still 16-times higher than the error rate that has been observed for the shorter reads generated by 454 and Illumina's MiSeq sequencing platforms. Although the longer reads frequently provided better classification, the wider adoption of this approach for 16S rRNA gene sequencing is likely limited by its high sequencing error and low yield of sequencing data relative to the other available platforms.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.778v1 | CC-BY 4.0 Open Access | rec: 7 Jan 2015, publ: 7 Jan 2015

Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system

Running title: 16S rRNA genes sequencing with PacBio

January 05, 2015

Authors: Patrick D. Schloss[1#], Sarah L. Westcott[1], Matthew L. Jenior[1], and Sarah K. Highlander[2]

* Correspondence: pschloss@umich.edu

1 Department of Microbiology and Immunology, 1500 W. Medical Center, University of Michigan, Ann Arbor, MI 48109

2 Genomic Medicine, J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA 92307

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.778v1 | CC-BY 4.0 Open Access | rec: 7 Jan 2015, publ: 7 Jan 2015

Abstract

Over the past 10 years, microbial ecologists have largely abandoned sequencing 16S rRNA genes by the Sanger sequencing method and have instead adopted highly parallelized sequencing platforms. These new platforms, such as 454 and Illumina's MiSeq, have allowed researchers to obtain millions of high quality, but short sequences. These platforms have allowed researchers to significantly improve the design of their experiments. The tradeoff has been the decline in the number of full-length reference sequences that are deposited into databases. To overcome this problem, we tested the ability of the PacBio Single Molecule, Real-Time (SMRT) DNA sequencing platform to generate sequence reads from the 16S rRNA gene. We generated sequencing data from the V4, V3-V5, V1-V3, V1-V6, and V1-V9 variable regions from within the 16S rRNA gene from a synthetic mock community and natural samples collected from human feces, mouse feces, and soil. The synthetic mock community allowed us to assess the actual sequencing error rate and how that error rate changed when different curation methods were applied. We developed a simple method based on sequence characteristics and quality scores to reduce the observed error rate for the V1-V9 region from 2.16% to 0.32%. Unfortunately, this error rate was still 16-times higher than the error rate that has been observed for the shorter reads generated by 454 and Illumina's MiSeq sequencing platforms. Although the longer reads frequently provided better classification, the wider adoption of this approach for 16S rRNA gene sequencing is likely limited by its high sequencing error and low yield of sequencing data relative to the other available platforms.

Keywords: Microbial ecology, bioinformatics, sequencing error

Introduction

Advances in sequencing technologies over the past 10 years have introduced considerable advances to the field of microbial ecology. Clone-based Sanger sequencing of the 16S rRNA gene has largely been replaced by various platforms produced by 454/Roche (e.g. Sogin et al. 2006), Illumina (e.g. Gloor et al. 2010), and IonTorrent (e.g. Junemann et al. 2012). It was once common to sequence fewer than 100 16S rRNA gene sequences from several samples using the Sanger approach (e.g. McCaig et al. 1999). Now it is common to generate thousands of sequences from each of several hundred samples (The Human Microbiome Consortium 2012). The advance in throughput has come at the cost of read length. Sanger sequencing regularly generated 800 nt per read and because the DNA was cloned, it was possible to obtain multiple reads per fragment to yield a full-length sequence from a representative single molecule. At approximately $8 (US) per sequencing read, most researchers have effectively decided that full-length sequences are not worth the increased cost relative to the cost of more recently developed approaches. There is still a clear need to generate high-throughput full-length sequence reads that are of sufficient quality that they can be used as references for analyses based on obtaining short sequence reads.

Historically, all sequencing platforms were created to primarily perform genome sequencing. When sequencing a genome, it is assumed that the same base of DNA will be sequenced multiple times and the consensus of multiple sequence reads is used to generate contigs. Thus, although an individual base call may have a high error rate, the consensus sequence will have a low error rate. To sequence the 16S rRNA gene researchers use conserved primers to amplify a sub-region from within the gene that is isolated from many organisms. Because the fragments are not cloned, it is not possible to obtain high sequence coverage from the same DNA molecule using these platforms. Thus, to reduce sequencing error rates it has become imperative to develop stringent sequence curation and denoising algorithms (Kozich et al. 2013; Schloss et al.

2011). There has been a tradeoff between read length, number of reads per sample, and the error rate. For instance, we recently demonstrated that using the Illumina MiSeq and the 454 Titanium platforms the raw error rate varies between 1 and 2% (Kozich et al. 2013; Schloss et al. 2011). Yet, it was possible to obtain error rates below 0.02% by adopting various denoising algorithms. However, the resulting fragments were only 250-nt long. In the case of 454 Titanium, extending the length of the fragment introduces length-based errors and in the case of the Illumina MiSeq, increasing the length of the fragment reduces the overlap between the read pairs reducing the ability of each read to mutually reduce the sequencing error. Inadequate denoising of sequencing reads can have many negative effects including limited ability to identify chimeras (Edgar et al. 2011; Haas et al. 2011) and inflation of alpha- and beta-diversity metrics (Huse et al. 2010; Kozich et al. 2013; Kunin et al. 2010; Schloss et al. 2011). Although MiSeq and 454 enjoy widespread use in the field, the MiSeq platform is emerging as the leader because of the ability to sequence 15-20 million fragments that can be distributed across hundreds of samples for less than $5000 (US).

As these sequencing platforms have grown in popularity, there has been a decline in the number of full-length 16S rRNA genes being deposited into GenBank that could serve as references. This is particularly frustrating since the technologies have significantly improved our ability to detect and identify novel populations for which we lack full-length reference sequences. A related problem is the perceived limitation that the short reads generated by the 454 and Illumina platforms cannot be reliably classified to the genus or species level. Previous investigators have utilized simulations to demonstrate that increased read lengths usually increase the accuracy and sensitivity of classification against reference databases (Liu et al. 2008; Wang et al. 2007; Werner et al. 2012). There is clearly a need to develop sequencing technologies that will allow researchers to generate high quality full-length 16S rRNA gene sequences in a high throughput manner.

New advances in single molecule sequencing technologies, such as the platform produced by Pacific Biosciences (PacBio), offer the opportunity to once again obtain full-length sequence reads with a high depth of coverage from a large number of samples. To this point, the PacBio Single Molecule, Real-Time (SMRT) DNA Sequencing System has received limited application in the microbial ecology research domain (Fichot & Norman 2013; Mosher et al. 2013; Mosher et al. 2014). The SMRT system ligates hairpin adapters (i.e. SMRTbells) to the ends of double-stranded DNA. Although the DNA molecule is linear, it is effectively circularized allowing the sequencing polymerase to process around the molecule multiple times (Au et al. 2012). According to Pacific Biosciences the platform is able to generate median read lengths longer than 8 kb with the P4-C2 chemistry; however, the single pass error rate is approximately 15%. Given the circular nature of the DNA fragment, the full read length can be used to cover the DNA fragment multiple times resulting in a reduced error rate. Therefore, one should be able to obtain multiple coverage of the full 16S rRNA gene at a reduced error rate.

Despite the opportunity to potentially generate high-quality full-length sequences, the Pacific Biosciences platform has not been widely adopted for sequencing 16S rRNA genes (Fichot & Norman 2013; Mosher et al. 2014). Previous studies utilizing the technology have removed reads with mismatched primers and barcodes, ambiguous base calls, and low quality scores (Fichot & Norman 2013). Others have utilized the platform without describing the bioinformatic pipeline that was utilized (Mosher et al. 2014). Regardless of the curation methods, the error rates associated with sequencing the 16S rRNA gene on the platform have never been reported. In the current study, we assessed the quality of data generated by the Pacific Biosciences sequencer and whether it could fill the need for generating high-quality, full-length sequence data. We hypothesized that by modulating the 16S rRNA gene fragment length we could alter the read depth and obtain reads longer than are currently available by the 454 and Illumina platforms but with the same quality. To test this hypothesis, we developed a sequence curation pipeline that was optimized by reducing the sequencing error rate of a mock bacterial community with known composition. The resulting pipeline was then applied to 16S rRNA gene fragments that were isolated from soil and human and mouse feces.

Materials and Methods

Community DNA. We utilized genomic DNA isolated from four communities. These same DNA extracts were previously used to develop an Illumina MiSeq-based sequencing strategy (Kozich et al. 2013). Briefly, we used a “Mock Community” composed of genomic DNA from 21 bacterial strains: *Acinetobacter baumannii* ATCC 17978, *Actinomyces odontolyticus* ATCC 17982, *Bacillus cereus* ATCC 10987, *Bacteroides vulgatus* ATCC 8482, *Clostridium beijerinckii* ATCC 51743, *Deinococcus radiodurans* ATCC 13939, *Enterococcus faecalis* ATCC 47077, *Escherichia coli* ATCC 70096, *Helicobacter pylori* ATCC 700392, *Lactobacillus gasseri* ATCC 33323, *Listeria monocytogenes* ATCC BAA-679, *Neisseria meningitidis* ATCC BAA-335, *Porphyromonas gingivalis* ATCC 33277, *Propionibacterium acnes* DSM 16379, *Pseudomonas aeruginosa* ATCC 47085, *Rhodobacter sphaeroides* ATCC 17023, *Staphylococcus aureus* ATCC BAA-1718, *Staphylococcus epidermidis* ATCC 12228, *Streptococcus agalactiae* ATCC BAA-611, *Streptococcus mutans* ATCC 700610, *Streptococcus pneumoniae* ATCC BAA-334. The mock community DNA is available through BEI resources (v3.1, HM-278D). Genomic DNAs from the three other communities were obtained using the MO BIO PowerSoil DNA extraction kit. The human and mouse fecal samples were obtained using protocols that were reviewed and approved by the University Committee on Use and Care of Animals (Protocol #PRO00004877) and the Institutional Review Board at the University of Michigan (Protocol #HUM00057066). The human stool donor provided informed consent.

Library generation and sequencing. The DNAs were each amplified in triplicate using barcoded primers targeting the V4, V1-V3, V3-V5, V1-V5, V1-V6, and V1-V9 variable regions (Table 1). The primers were synthesized so that the 5’ end of the forward and reverse primers were each tagged with a 5-nt barcode sequence to allow multiplexing of samples within a single sequencing run. Methods describing PCR, amplicon cleanup, and pooling were described previously (Kozich et al. 2013). The SMRTbell adapters were ligated onto the PCR products and the libraries were sequenced at the University of Michigan DNA Sequencing Core using the P4-C2 chemistry on a PacBio RS II SMRT DNA Sequencing System.

Data analysis. All sequencing data were curated using mothur (Schloss et al. 2009) and analyzed using the R programming language (R Core Team 2014). The raw data can be obtained from the Sequence Read Archive at NCBI under accession SRP051686, which are associated with BioProject PRJNA271568. Several specific features were incorporated into mothur to facilitate the analysis of PacBio sequence data. First, because non-ambiguous base calls are assigned to Phred quality scores of zero, the consensus fastq files were parsed so that scores of zero were interpreted as corresponding to an ambiguous base call (i.e. N) in the fastq.info command using the pacbio=T option. Second, because the consensus sequence can be generated in the forward and reverse complement orientations, a checkorient option was added to the trim.seqs command in order to identify the proper orientation. These features were incorporated into mothur v.1.30. Because chimeric molecules can be generated during PCR and would artificially inflate the sequencing error, it was necessary to remove these data prior to assessing the error rate. Because we knew the true sequences for the strains in the mock community we could calculate all possible chimeras between strains in the mock community (*in silico* chimeras). If a sequence read was 3 or more nucleotides more similar to an *in silico* chimera than it was to a non-chimeric reference sequence, it was classified as a chimera and removed from further consideration. Identification of *in silico* chimeras and calculation of sequencing error rates was performed using the seq.error command in mothur (Schloss et al. 2011). *De novo* chimera detection was also performed on the mock and other sequence data using the abundance-based algorithm implemented in UCHIME (Edgar et al. 2011). Sequences sequences were aligned against a SILVA-based reference alignment (Pruesse et al. 2007) using a profile-based aligner (Schloss 2009) and were classified against the SILVA (Pruesse et al.

2007), RDP (Cole et al. 2014), and greengenes (Werner et al. 2012) reference taxonomies using a naive Bayesian classifier (Wang et al. 2007). Sequences were assigned to operational taxonomic units using the average neighbor clustering algorithm with a 3% distance threshold (Schloss & Westcott 2011). Detailed methods including this paper as an R markdown file are available as a public online repository (http://github.com/SchlossLab/Schloss_PacBio16S_PeerJ_2015).

Results and Discussion

The PacBio error profile. To build a sequence curation pipeline, we first needed to characterize the error rate associated with sequencing the 16S rRNA gene. We observed an average sequencing error rate of 1.80%. Insertions, deletions, substitutions, and ambiguous base calls accounted for 45.3, 17.3, 35.8, and 2.1% of the errors, respectively. The substitution errors were equally likely and all four bases were equally likely to cause insertion errors. Interestingly, guanines (44.6%) and cytosines (34.5%) were more likely to be deleted than adenines (11.4%) or thymidines (9.5%). When we considered the Phred quality score of each base call, we observed a median quality score of 72 for correct base calls and scores of 22 and 20 for substitutions and insertions, respectively (Figure 1A). Although there was a broad distribution of quality scores with each type of base call, the errors could largely be distinguished from the correct base calls.

A basic sequence curation procedure. To establish a simple curation procedure, we culled any sequence that contained an ambiguous base call, had a string of the same base repeated 9 or more times, did not start and end at the expected alignment coordinates for that region of the 16S rRNA gene, or that was chimeric. This reduced the experiment-wide error rate from 1.80 to 0.90%. This basic procedure resulted in the removal of between 4.0 (V4) and 32.2 (V1-V9)% of the reads. The percentage of reads removed increased with the length of the fragment (Figure 2). The number of reads removed because of the presence of ambiguous base calls was similar to the number of reads that were removed for not fully aligning to the correct region within the 16S rRNA gene (Table 2). The latter class of errors was generally due to sequence truncations that could not be explained.

Identifying correlates of increased sequencing error. In contrast to the 454 and Illumina-based platforms where the sequencing quality decays with length, the consensus sequencing approach employed by the PacBio sequencer is thought to generate a uniform distribution of errors. This makes it impossible to simply trim sequences to high quality regions. Therefore, we sought to identify characteristics within sequences that would allow us to identify and remove those sequences with errors using three different approaches. First, we hypothesized that errors in the barcode and primer would be correlated with the error rate for the entire sequence. We observed a strong relationship between the number of mismatches to the barcodes and primers and the error rate of the rest of the sequence fragment (Figure 1B). Although allowing no mismatches to the barcodes and primers yielded the lowest error rate, that stringent criterion removed a large fraction of the reads from the dataset and allowing at most one mismatch marginally increased the error rate while preserving more sequences in the dataset (Figure 2). Second, we hypothesized that increased sequencing coverage should yield lower error rates. We found that once we had obtained 10-fold coverage of the fragments, the error rate did not change appreciably (Figure 1C). When we compared the error rates of reads with at least 10-fold coverage to those with less coverage, we reduced the error rate by 26.5 to 29.7% for each region except the V4 region for which the error rate was reduced by 53%. Third, based on the earlier analysis associating errors with quality scores, we used two quality score-based approaches for identifying reads with errors (Figure 3). We calculated the minimum average quality score across all 50-nt windows within each sequence and we also calculated the average quality score across each sequence. We then associated both methods of calculating the average quality score with the error rate of the reads and the fraction of sequences that would be retained if each threshold were selected. Using the sliding window approach we did not observe any clear break points indicating that one quality score would be better than another (Figure 3AB). In contrast, using the whole sequence quality score average we observed a decrease in the error rate and the fraction of sequences retained when the threshold was increased above 60 (Figure 3CD). When we used this threshold, we were able to reduce the error rate by 32.8 to 56.1% (Figure 2A). We noted that the fraction of reads retained decreased as the length of the fragment increased with retention of 86.9% of the V4 reads and 50.1% of the V1-V9 reads (Figure 2B). Next, we asked whether which combinations of culling reads with mismatches to the expected barcodes and primers, less than 10-fold sequencing coverage, and an average quality score less than 60 made the most meaningful reductions in the error rate while preserving the most reads when implemented with the basic curation pipeline (Figure 2B). We observed similar error rates when we required one or fewer mismatches to the barcodes and primers and an average quality score above 60 as when we also required a minimum 10-fold coverage. Culling sequences that had more than one mismatch to the barcodes and primers and those with an average quality score less than 60 reduced the error rate to between 0.22 and 0.97. This procedure resulted in the removal of 18 and 53% of the reads (Figure 2). The remainder of this paper uses this sequence curation approach.

Pre-clustering sequences to further reduce sequencing noise. Previously, we implemented a pre-clustering algorithm where sequences were sorted by their abundance in decreasing order and rare sequences are clustered with a more abundant sequence if the rare sequences have fewer mismatches than a defined threshold when compared to the more abundant sequence. The recommended threshold was a 1-nt difference per 100-nt of sequence data. For example, the threshold for 250 bp fragment from the V4 region would be 2 nt or 14 for the 1458 bp V1-V9 fragments. This approach removes residual PCR and sequencing errors while not overwhelming the resolution needed to identify OTUs that are based on a 3% distance threshold. The tradeoff of this approach is that one would unable to differentiate V1-V9 sequences that truly differed by less than 14 nt. When we applied this approach to our PacBio data, we observed a reduction in the error rate between 15 (V1-V3 and V3-V5) and 44% (V1-V5). The final error rates varied between 0.14 (V4) and 0.83% (V3-V5); the full-length, V1-V9, fragments had an error rate of 0.32% (Figure 2B). These error rates are 7-40 times higher than what we have previously observed using the 454 and Illumina MiSeq platforms (0.02%)(Kozich et al. 2013; Schloss et al. 2011)

Effects of error rates on OTU assignments. The sequencing error rate is known to affect the number of OTUs that are observed (Schloss et al. 2011). For each region, we determined that if there were no chimeras or PCR or sequencing errors, then we would expect to find 20 OTUs. When achieved perfect chimera removal, but allowed for PCR and sequencing errors, we observed between 6 (V4) and 63.1 (V3-V5) extra OTUs. The range in the number of extra OTUs was largely explained by the sequencing error rate (Pearson's R=0.91). Next, we determined the number of OTUs that were observed when we used UCHIME to identify chimeric sequence. Under these more realistic conditions, we observed between 7.4 (V4) and 86.8 (V3-V5) extra OTUs. Finally, we calculated the number of OTUs in the soil, mouse, and human samples using the same pipeline with chimera detection and removal based on the UCHIME algorithm. Again, we found that there was a strong correlation between the number of observed OTUs and the error rate for the soil (R=0.62), mouse (R=0.90), and human samples (R=0.72). These results underscore the effect of sequencing error on the inflation of the number of observed OTUs.

Increasing sequence length improves classification. We classified all of the sequence data we generated using the naïve Bayesian classifier using the RDP, SILVA, and greengenes reference taxonomies (Figure 4). In general, increasing the length of the region improved the ability to assign the sequence to a genus or species. Interestingly, each of the samples we analyzed varied in the ability to assign its sequences to the depth of genus or species. Furthermore, the reference database that did the best job of classifying the sequences varied by sample type. For example, the SILVA reference did the best for the human feces and soil samples and the RDP did the best for the mouse feces samples. An advantage of the greengenes database is that it contains information for 2,514 species-level lineages for 11% of the reference sequences; the other databases only provided taxonomic data to the genus level. There was a modest association between the length of the fragment and the ability to classify sequences to the species-level for the human samples; there was no such association for the mouse and soil samples. In fact, at most 4.0% of the soil sequences and 3.8% of the mouse sequences could be classified to a species. These results indicate that the ability to classify sequences to the genus or species level is a function of read length, sample type, and the reference database.

Sequencing errors are not random. Above, we described that although there was no obvious bias in the substitution or insertion rate, we did observe that guanines and cytosines were more likely to be deleted than adenines and thymidines. This lack of randomness in the error profile suggested that there might be a systematic non-random distribution of the errors across the sequences. This would manifest itself by the creation of duplicate sequences with the same error. Because we were able to obtain a large number of reads from the mock communities where we sequenced the V4 (N=17361), V1-V5 (N=8061 sequences), and V3-V5 (N=4854) regions, we investigated the mock community data from these regions further. We identified all of the sequences that had a 1-nt difference to the true sequence. For these three regions, a majority of the sequences with 1-nt errors were only observed once (V4: 75.6%, V1-V5: 82.8%, V3-V5: 79.8%). We found that the frequency of the most abundant 1-nt error paralleled the number of sequences. There were two sequences in the V4 dataset that occurred 76 times, one sequence in the V1-V5 dataset that occurred 30 times, and one sequence in the V3-V5 dataset that occurred 17 times. Contrary to previous reports (Carneiro et al. 2012; Koren et al. 2012), these results indicate that reproducible errors occur with the PacBio sequencing platform and that they can be quite frequent.

Conclusions

The various sequencing platforms that are available to microbial ecologists are able to fill unique needs and have their own strengths and weaknesses. For sequencing the 16S rRNA gene, the 454 platform is able to generate a moderate number of high-quality 500-nt sequence fragments (error rates below 0.02%) (Schloss et al. 2011) and the MiSeq platform is able to generate a large number of high-quality 250-nt sequence fragments (error rates below 0.02%) (Kozich et al. 2013). The promise of the PacBio sequencing platform was the generation of high-quality near full-length sequence fragments. As we have shown in this study, it is possible to generate near full-length sequences; however, the error rate associated with those reads is considerable (i.e. 0.32%) and requires a level of sequencing coverage that is not commonly observed in a typical sequencing run. This results in the generation of a small number of low quality full-length sequences. When we considered the shorter V4 region, which is similar in length to what is sequenced by the MiSeq platform, the error rates we observed with the PacBio platform were nearly 5-fold higher than what has previously been reported. It appears that the promise offered by the PacBio platform has not been realized.

The widespread adoption of the 454 and MiSeq platforms and decrease in the use of Sanger sequencing for the 16S rRNA gene has resulted in a decrease in the generation of the full-length reference sequences that are needed for performing phylogenetic analyses and designing lineage specific PCR primers and fluorescent *in situ* hybridization (FISH) probes. It remains to be determined whether the elevated error rates we observed for full-length sequences are prohibitive for these applications. We can estimate the distribution of errors assuming that the errors follow a binomial distribution along the length of the 1,500 nt gene with the error rate that we achieved from the V1-V9 mock community data prior to pre-clustering the sequences, which was 0.52% (Figure 5). Under these conditions one would only expect 0.04% of the sequences to have no errors. In fact, 95% of the reads would have at least 3 errors and 50% of the reads would have at least 8 errors. If the error rate could be dropped to 0.25%, then 95% of the reads would have at least 1 error and 50% of the reads would have at least 4 errors. If it were possible to replicate the low error rates we have previously observed using the 454 and Illumina MiSeq platforms, which was 0.02%, then we would expect 74.1% of the sequences to have no errors. In fact, 95% of the reads would have 1 or fewer errors. Although full-length sequence data is highly desired, at this point, it does not appear that the PacBio platform can provide the data of sufficient quality to fill the niche of generating reference sequences.

Full-length sequences are frequently seen as a panacea to overcome the limitations of taxonomic classifications. The ability to classify each of our sample types benefited from the generation of full-length sequences. It was interesting that the benefit varied by sample type and database. For example, using the mouse libraries, the ability to classify each of the regions differed by less than 5% when classifying against the SILVA and greengenes databases. The effect of the database that was used was also interesting. The RDP database outperformed the other databases for the mouse samples and the SILVA database outperformed the others for the human and soil samples. The three databases were equally effective for classifying the mock community. Finally, since only the greengenes database provided species-level information for its reference sequences it was the only database that allowed for resolution of species-level classification. The sequences from the mouse and soil libraries were not effectively classified to the species level (all less than 10%). In contrast, classification of the human libraries resulted in more than 40% of the sequences being classified to a genus, regardless of the region. That the variation in species-level classification for the human libraries was less than 10% suggests that the benefit of added length is minimal considering the lower sequencing yield.

The development of newer sequencing technologies continue to advance and there is justifiable excitement to apply these technologies to sequence the 16S rRNA gene. Although it is clearly possible to generate sequencing data from these various platforms, it is critical that we assess the platforms for their ability to generate high quality data and the particular niche that the new approach will fill. With this in mind, it is essential that researchers utilize mock communities as part of their experimental design so that they can quantify their error rates. The ability to generate large numbers of near full-length 16S rRNA gene sequences is an exciting advance. At this point, the excitement must be tempered by the appreciation that the error rates limit the application of the approach.

Acknowledgements

The Genomic DNA from Microbial Mock Community A (Even, Low Concentration, v3.1, HM-278D) was obtained through the NIH Biodefense and Emerging Infections Research Resources Repository, NIAID, NIH as part of the Human Microbiome Project.

Funding statement

This study was supported by grants from the NIH (R01HG005975, R01GM099514 and P30DK034933 to PDS and U54HG004973 to SKH).

References

Au KF, Underwood JG, Lee L, and Wong WH. 2012. Improving PacBio long read accuracy by short read alignment. *PLoS ONE* 7:e46679.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, and DePristo MA. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* 13:375.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, and Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:D633-642.

Edgar RC, Haas BJ, Clemente JC, Quince C, and Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194-2200.

Fichot EB, and Norman RS. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *MIcrobiome* 1:10.

Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R, and Reid G. 2010. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS ONE* 5:e15406.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methe B, DeSantis TZ, Petrosino JF, Knight R, and Birren BW. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 21:494-504.

Huse S, Mark Welch D, Morrison H, and Sogin M. 2010. Ironing out the wrinkles in the rare biosphere. *Environ Microbiol* In press.

Junemann S, Prior K, Szczepanowski R, Harks I, Ehmke B, Goesmann A, Stoye J, and Harmsen D. 2012. Bacterial community shift in treated periodontitis patients revealed by Ion Torrent 16S rRNA gene amplicon sequencing. *PLoS ONE* 7:e41606.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, and Adam MP. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* 30:693-700.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, and Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 79:5112-5120.

Kunin V, Engelbrektson A, Ochman H, and Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 12:118-123.

Liu Z, DeSantis TZ, Andersen GL, and Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 36:e120.

McCaig AE, Glover LA, and Prosser JI. 1999. Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl Environ Microbiol* 65:1721-1730.

Mosher JJ, Bernberg EL, Shevchenko O, Kan J, and Kaplan LA. 2013. Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S rRNA genes from environmental samples. *J Microbiol Methods* 95:175-181.

Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J, and Kaplan LA. 2014. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *J Microbiol Methods* 104:59-60.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, and Glockner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188-7196.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Schloss PD. 2009. A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS ONE* 4:e8230.

Schloss PD, Gevers D, and Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310.

Schloss PD, and Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol* 77:3219-3226.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, and Weber CF. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537-7541.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, and Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 103:12115-12120.

The Human Microbiome Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207-214.

Wang Q, Garrity GM, Tiedje JM, and Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261-5267.

Werner JJ, Koren O, Hugenholtz P, Desantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, and Ley RE. 2012. Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. *ISME J* 6:94-103.

Figure 1(on next page)

Summary of errors in data generated using PacBio sequencing platform to sequence various regions within the 16S rRNA gene.

Quality scores varied with error types (A). The sequencing error rate of the amplified gene fragments increased with mismatches to the barcodes and primers (B). The sequencing error rate declined with increasd sequencing coverage; however, increasing the sequencing depth beyond 10-fold coverage had no meaningful effect on the sequencing error rate (C).

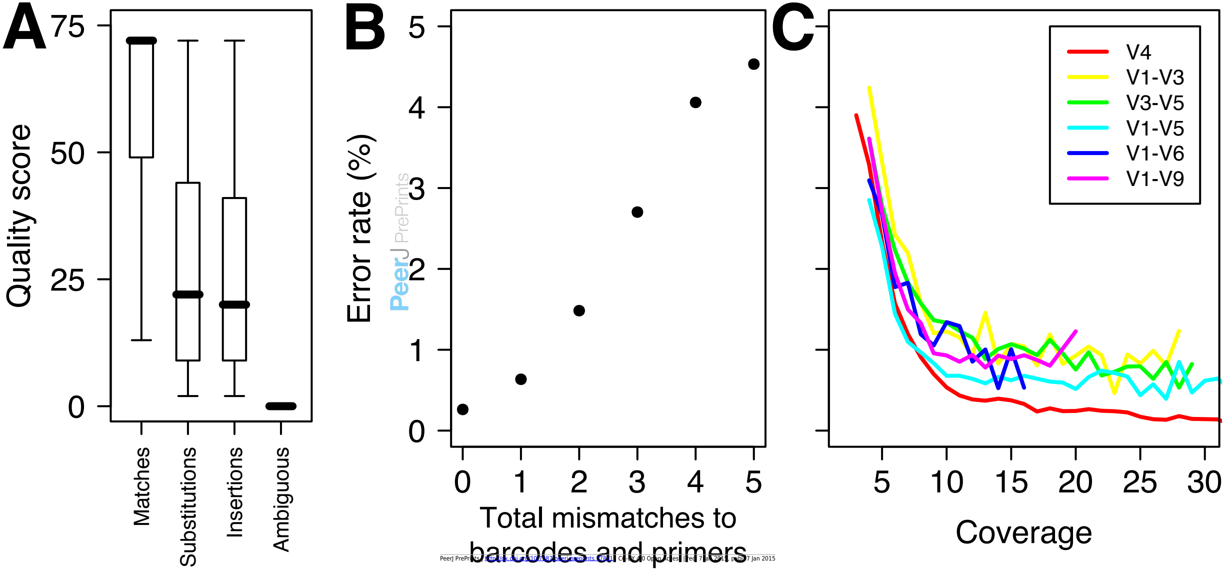

A
75
50
25
0
Quality score
Matches
Substitutions
Insertions
Ambiguous
B
0
1
2
3
4
5
Error rate (%)
0
1
2
3
4
5
Total mismatches to
barcodes and primers
C
V4
V1-V3
V3-V5
V1-V5
V1-V6
V1-V9
5
10
15
20
25
30
Coverage

Figure 2(on next page)

Change in error rate (A) and the percentage of sequences that were retained (B) when using various sequence curation methods.

The condition that was used for downstream analyses is indicated by the star. The plotted numbers represent the region that was sequenced. For example "15" represents the data for the V1-V5 region.

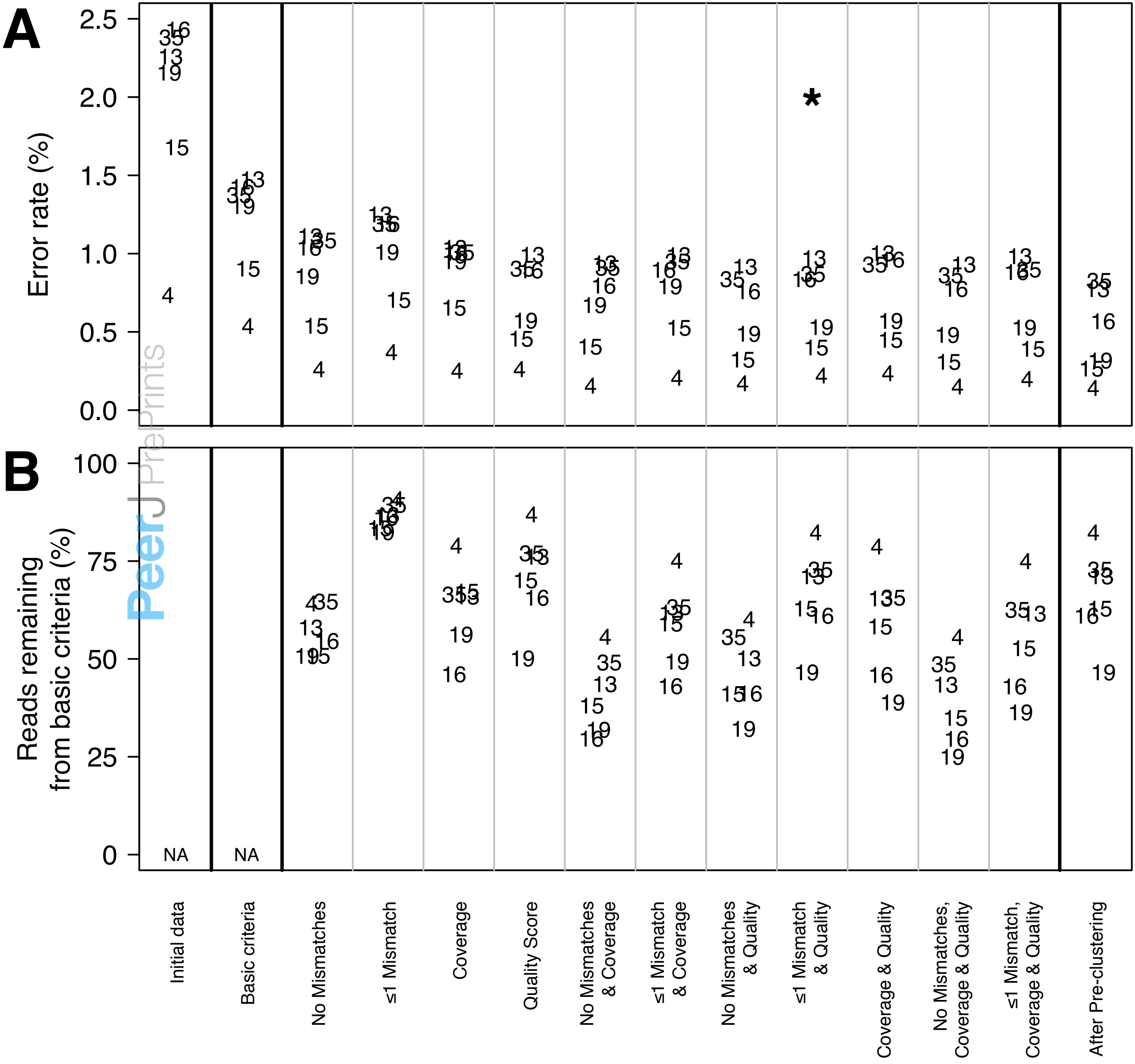

A
B
Error rate (%)
2.5
2.0
1.5
1.0
0.5
0.0
*
Reads remaining from basic criteria (%)
100
75
50
25
0
NA
NA
Initial data
Basic criteria
No Mismatches
≤1 Mismatch
Coverage
Quality Score
No Mismatches & Coverage
≤1 Mismatch & Coverage
No Mismatches & Quality
≤1 Mismatch & Quality
Coverage & Quality
No Mismatches, Coverage & Quality
≤1 Mismatch, Coverage & Quality
After Pre-clustering

Figure 3(on next page)

The relationship between the error rate of each region and the composite quality scores for the sequences.

The error rates (A and C) and percentage of sequences (B and D) were calculated for the reads that had a composite quality score above the plotted value. The composite quality scores were calculated by either determining the minimum value of the average quality score wihin all 50-nt windows within each region (A and B) or by calculating the average quality score across the entire sequence read (C and D).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.778v1 | CC-BY 4.0 Open Access | rec: 7 Jan 2015, publ: 7 Jan 2015

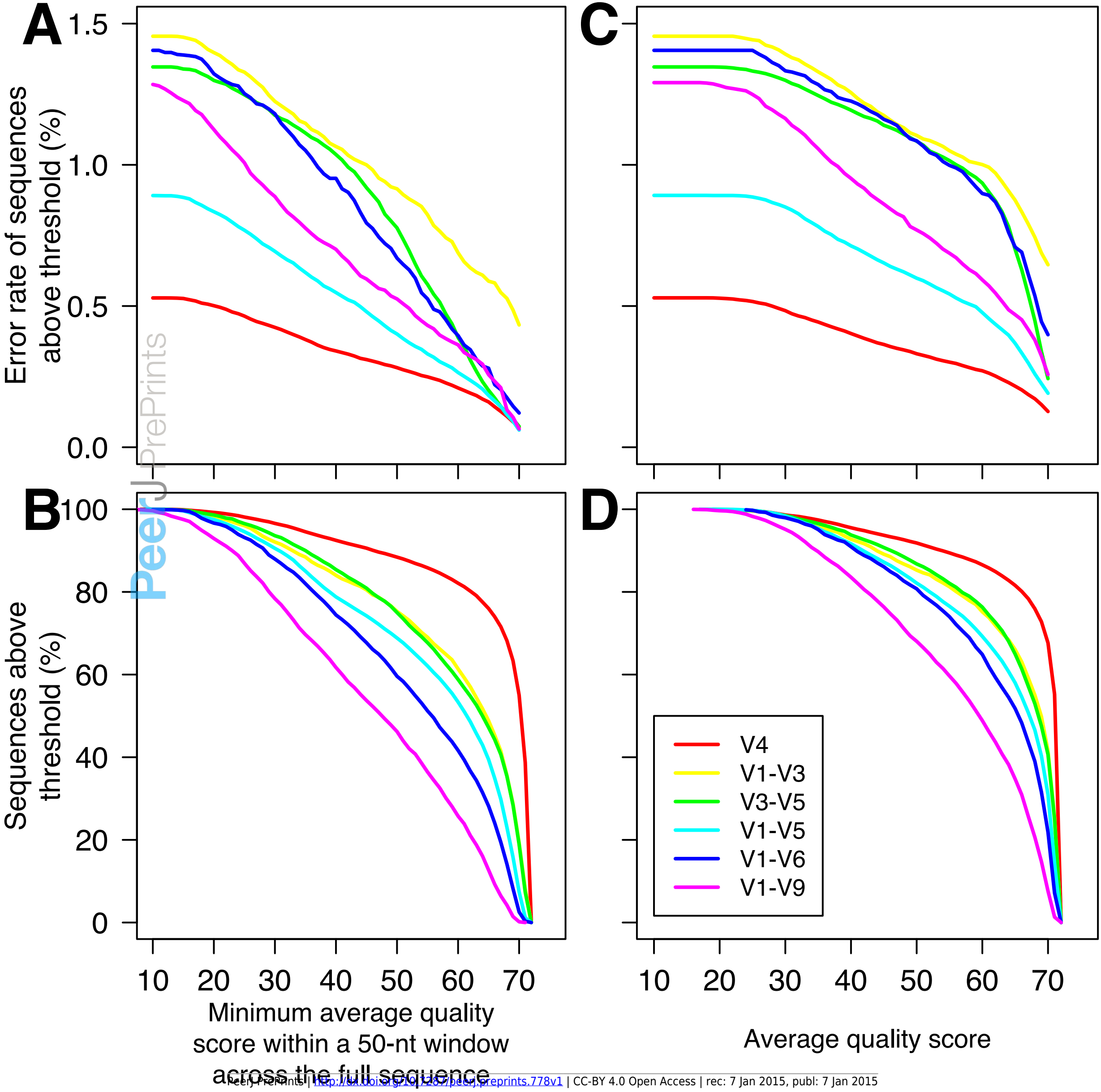

A
B
C
D
Error rate of sequences above threshold (%)
0.0
0.5
1.0
1.5
Sequences above threshold (%)
0
20
40
60
80
100
10
20
30
40
50
60
70
Minimum average quality score within a 50-nt window across the full sequence
Average quality score
V4
V1-V3
V3-V5
V1-V5
V1-V6
V1-V9

Figure 4(on next page)

Percentage of unique sequences that could be classified.

Classifications were performed using taxonomy references curated from the RDP, SILVA, or greengenes databases for the four types of samples that were sequenced across the six regions from the 16S rRNA gene. Only the greengenes taxonomy reference provided species-level information.

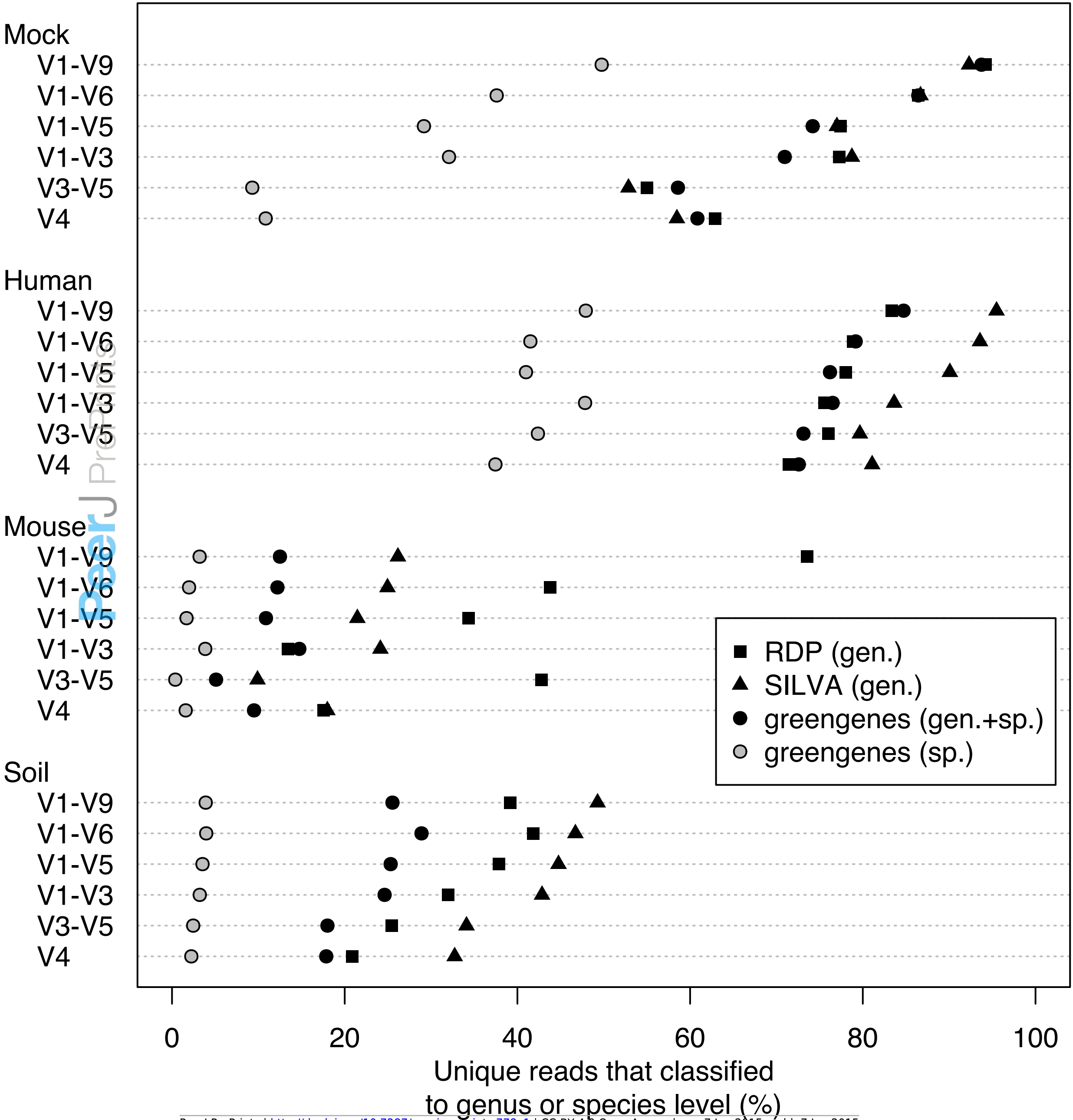
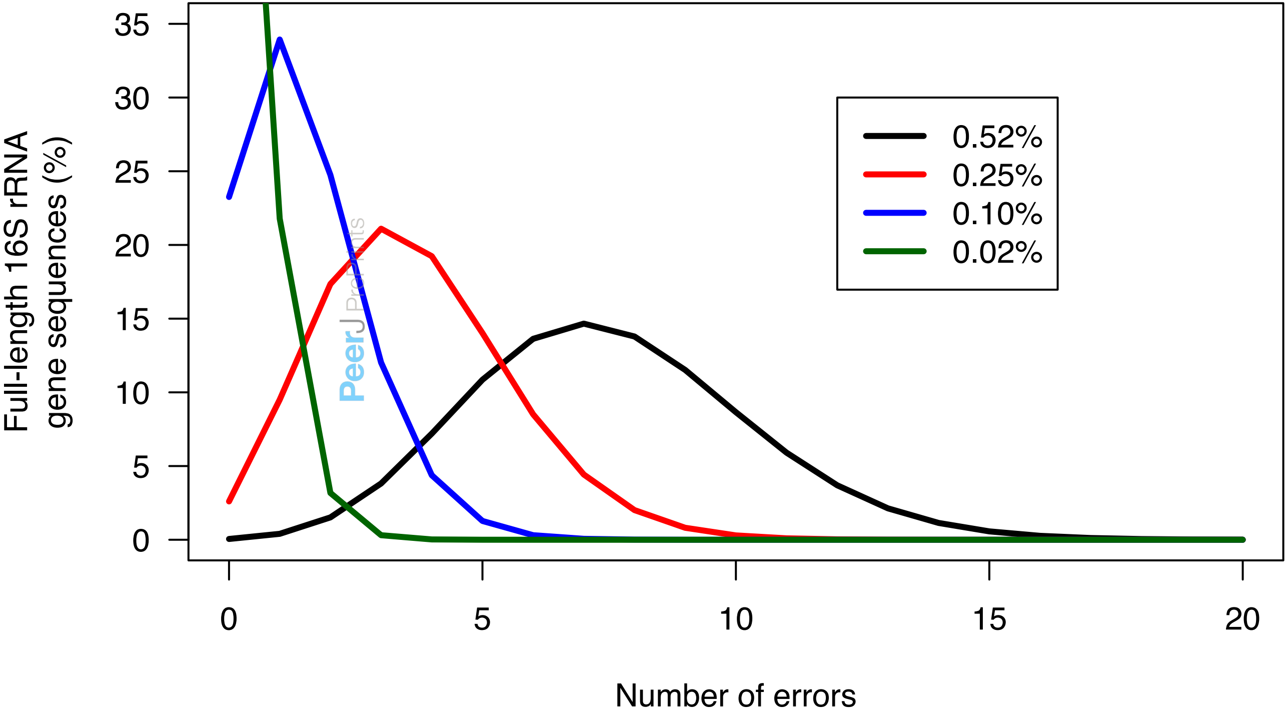

Mock
V1-V9
V1-V6
V1-V5
V1-V3
V3-V5
V4
Human
V1-V9
V1-V6
V1-V5
V1-V3
V3-V5
V4
Mouse
V1-V9
V1-V6
V1-V5
V1-V3
V3-V5
V4
Soil
V1-V9
V1-V6
V1-V5
V1-V3
V3-V5
V4
■ RDP (gen.)
▲ SILVA (gen.)
● greengenes (gen.+sp.)
○ greengenes (sp.)
0
20
40
60
80
100
Unique reads that classified
to genus or species level (%)

Figure 5(on next page)

The percentage of V1-V9 sequences that were predicted to have between 0 and 20 errors as a function of the error rate of the sequences.

The highest error rate, 0.52%, corresponds to what was observed before the pre-clustering step. The smallest error rate (0.02%) corresponds to our previous observations using the 454 and MiSeq sequencing platforms. The predicted number of errors was assumed to follow a binomial distribution.



35
30
25
20
15
10
5
0
Full-length 16S rRNA gene sequences (%)
0
5
10
15
20
Number of errors
0.52%
0.25%
0.10%
0.02%

Table 1(on next page)

Tables

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.778v1 | CC-BY 4.0 Open Access | rec: 7 Jan 2015, publ: 7 Jan 2015

Table 1. Summary of the primer pairs used to generate the 16S rRNA gene fragment fragments and the characteristics of each region.

	Forward	Reverse	*E. coli* coordinates[a]	Length (bp)[b]
V4	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	515-806	253
V3-V5	CCTACGGGAGGCAGCAG	CCCGTCAATTCMTTTRAGT	341-927	551
V1-V3	AGRGTTTGATYMTGGCTCAG	ATTACCGCGGCTGCTGG	8-534	490
V1-V5	AGRGTTTGATYMTGGCTCAG	CCCGTCAATTCMTTTRAGT	8-927	881
V1-V6	AGRGTTTGATYMTGGCTCAG	ACRACACGAGCTGACGAC	8-1078	1033
V1-V9	AGRGTTTGATYMTGGCTCAG	GGYTACCTTGTTACGACTT	8-1510	1464

[a] The coordinates where the start and end of the forward and reverse primers anneal, respectively.

[b] The number of bases between the primers.

Table 2. Summary of the reasons that sequences were excluded because of the basic sequence curation steps

	Initial sequences (N)	Good reads (%)	Wrong start/end position (%)	Excessively long homopolymers (%)	Ambiguous base calls (%)	Sequences remaining (N)
V4	21841	96.0	2.9	0.1	1.5	20974
V3-V5	5212	84.0	10.0	0.1	7.5	4378
V1-V3	7236	77.3	15.6	0.2	11.0	5594
V1-V5	14875	79.1	11.5	0.2	12.5	11764
V1-V6	2220	72.6	11.4	0.1	19.4	1611
V1-V9	5003	67.8	18.0	0.5	17.5	3393