Short read lengths recover ecological patterns in 16S rRNA gene amplicon data

Stephanie Jurburg[1]

[1]Helmholtz-Centre for Environmental Research - UFZ

August 29, 2023

Abstract

Metabarcoding is an increasingly popular and accessible method for assessing bacterial communities across a wide range of environments, and as the sequence data archives grow, sequence data reuse will likely become an important source of novel insights into the ecology of microbes. While literature on the benefits of longer read lengths for the study of microbial communities, little is known about the (re)usability of shorter (<200 bp) read lengths, but this information is essential to improve the reuse and comparability of metabarcoding data across studies. This study reanalyzed three 16S rRNA datasets targeting aquatic, animal-associated, and soil microbiomes, and evaluated how processing the sequence data across a range of read lengths affected the resulting taxonomic assignments, biodiversity metrics, and differential (i.e., before-after treatment) analyses. Short read lengths successfully recovered ecological patterns, and limited increases in resolution were observed beyond 100 bp reads across environments. Furthermore, abundance-weighted diversity metrics (e.g., Inverse Simpson index or Bray-Curtis dissimilarities) were more robust to variation in read lengths. Importantly, the total number of ASVs detected increased with read length, highlighting the need to consider metabarcoding-derived diversity estimates within the context of the bioinformatics parameters selected. This study provides evidence-based guidelines for the processing of short reads.

Introduction

The 16S rRNA gene is about 1,550 bp long and encodes the small subunit ribosomal RNA molecules of ribosomes. Originally used by Woese and Fox to examine the phylogeny of prokaryotes, (Woese & Fox, 1977), the 16S rRNA gene currently serves as a molecular clock (Woese, 1987) and as a means for differentiating prokaryotic taxa (Benjamin J Callahan, McMurdie, & Holmes, 2017; Clarridge, 2004). Sequencing this gene has revolutionized microbial ecology, allowing for the identification of microbes that cannot be studied in isolation, or exist in complex mixtures, and revealing the astounding complexity and ubiquity of microbes globally (Thompson et al., 2017).

The structure of the 16S rRNA transcript and its essential function in protein synthesis have limited the rate of evolutionary change in the gene, resulting in highly conserved regions that can be leveraged as primer targets that contain variable regions for sequencing (Clarridge, 2004). Initial sequence-based assessments of prokaryotic diversity relied on Sanger sequencing, which could sequence the entirety of the 16S rRNA gene of few reads at a high cost and effort. The advent of next generation sequencing technologies (e.g., Illumina HiSeq, IonTorrent), heretofore amplicon sequencing, allowed for the sequencing of a much greater number of sequences, but at a length of <600 base pairs (Caporaso et al., 2011).

Despite the development of third-generation sequencing techniques that allow the sequencing of the full-length of 16S rRNA genes and provide a higher taxonomic resolution (Johnson et al., 2019; Matsuo et al., 2021), amplicon sequencing of shorter segments remains the most accessible method for the identification of microbial communities. Amplicon sequencing data continues to grow exponentially in sequence archives (Jurburg, Konzack, Eisenhauer, & Heintz-Buschart, 2020), representing an important data resource for

future research, and has already provided important insights into the abundance of prokaryotes (e.g., (Louca, Mazel, Doebeli, & Parfrey, 2019)). A wealth of literature describes the limitations and biases of amplicon sequencing, including the impact of amplification (Brooks et al., 2015; Schloss, Gevers, & Westcott, 2011), bioinformatics processing (Kang, Deng, Crielaard, & Brandt, 2021; Marizzoni et al., 2020; Prodan et al., 2020), and hypervariable region (Bukin et al., 2019; Tremblay et al., 2015; Yang, Wang, & Qian, 2016; Yu, García-González, Schanbacher, & Morrison, 2008) on the resulting microbial diversity data.

Critically, while the positive impact of full 16S rRNA gene sequences on taxonomic assignment has been well documented (Curry et al., 2022; Johnson et al., 2019), the extent to which short read lengths (i.e., <200 base pairs) are able to recover higher-level taxonomic assignments and ecological patterns has received little attention. Understanding the opportunities and limitations of shorter 16S rRNA gene read lengths is essential, especially for the reuse of rapidly growing sequence data archives (Jurburg et al., 2020). Read truncation is a common practice that removes lower-quality read ends (e.g., (Ben J Callahan, Sankaran, Fukuyama, McMurdie, & Holmes, 2016)). Most bioinformatics workflows aim to maximize read length, however, allowing shorter read lengths can improve the comparability of sequences across datasets, allowing for re-analyses that target the identical 16S rRNA gene region and avoid biases that emerge from sequencing different, but overlapping target regions (Bukin et al., 2019; Tremblay et al., 2015; Yang et al., 2016; Yu et al., 2008), or from differential read lengths. Characterizing the impact of shorter read lengths on 16S rRNA gene-based ecological assessments may also serve for the integration of data from diverse platforms that produce a range of sequence lengths (e.g., from full gene sequencing with Nanopore to 150 bp with single-ended sequencing in HiSeq).

To examine the effect of sequence length on microbial diversity estimates, three datasets from disturbed soil, water, and animal microbiomes sequenced using the same primer set and sequencing platform across a gradient of read lengths were reprocessed. It was hypothesized that 1) shorter reads would result in a higher percentage of unclassified ASVs and 2) lower richness estimates, but that 3) the relationship between disturbed and undisturbed samples in each environment would still be detectable.

Materials and Methods

Reused data

To examine the effect of read length on diversity estimates, publicly available data from three experimental datasets were selected. These targeted the soil, water, and animal microbiomes before and 4 days after a single, strong disturbance. Each dataset had at least 5 replicates per time point, had sampled the system before, one day after, and 4 days after disturbance, and had sequenced the resulting DNA by targeting the 515-806 region using an Illumina MiSeq. This 16S rRNA region was selected due to its popularity (Jurburg et al., 2020), standardization (Thompson et al., 2017), and sensitivity (Zhang et al., 2023).

Soil microbiome data was obtained from a microcosm experiment (Jurburg et al., 2017), in which soils were exposed to microwave radiation, and destructively sampled over time. DNA was extracted using the MoBio PowerSoil DNA Extraction Kit (MoBio Laboratories, Carlsbad, CA, U.S.A.), and the resulting data are publicly available in NCBI's SRA under accession number PRJNA329541. Five replicate samples per time point for the treatment exposed to a single microwave disturbance, with samples taken before, one day after, and 4 days after the disturbance were selected. Where more replicates were available, the five largest files for each *dataset x time* combination were selected.

Aquatic microbiome data was obtained from a marine microcosm experiment exposing aquatic microbiota to cadmium and phenanthrene (Qian, Ding, Guo, Zhang, & Wang, 2017). Sampling was repeated within each microcosm over time. Water was filtered and DNA was extracted using the MoBio PowerSoil DNA Extraction Kit (MoBio Laboratories, Carlsbad, CA, U.S.A.), and the resulting data are publicly available in NCBI's SRA under accession number PRJDB4992. Samples simultaneously exposed to phenanthrene and cadmium, taken before, one day after, and 4 days after disturbance were selected, and all five replicates per time point were included.

Finally, animal microbiome data was obtained from an animal experiment in which experimental pigs were exposed to antibiotic treatments and their fecal microbiome was monitored daily (Jurburg, Cornelissen, de Boer, Smits, & Rebel, 2019). Fecal samples were snap-frozen, and DNA was extracted using the AGOWA mag Mini DNA Isolation Kit (AGOWA, Berlin, Germany), and the resulting data are publicly available in NCBI's SRA under accession number PRJNA528235. Animals exposed to clindamycin, and sampled before, one day after, and 4 days after disturbance were selected, and all five replicates per time point were included.

Sequence processing and data analysis

Sequence data and metadata were downloaded from NCBI and processed using the popular *dada2* pipeline (B J Callahan et al., 2016) and standard parameters (maxN=0; maxEE=2, truncQ=2). As our goal was to explore the impact of shorter read lengths on the taxonomic assignment of prokaryotes, and its impact on the ecological conclusions derived from the data, only forward read lengths from each dataset were selected. Importantly for sequence data reuse, reverse reads are often not available in archived sequence data (Jurburg et al., 2020), either because pair-ended sequencing was not performed or the reverse reads are not archived. Indeed, one of the datasets used (Qian et al., 2017) had merged paired ends prior to archiving. For each sample, read length was varied from 50-200 bp in intervals of 10 bp. This range of read lengths was selected as it represents the minimum output of all next generation sequencing technologies. Taxonomy was assigned using SILVA v138 (Quast et al., 2013). For all samples, the number of unassigned reads at each taxonomic level, and the percentage of original reads included in the final ASV table was recorded.

ASV tables were analyzed using *phyloseq (McMurdie & Holmes, 2013)* and *vegan (Oksanen et al., 2007)* . To compare diversity estimates, all versions of each dataset were rarefied to the lowest number of reads (23,354 reads for the water dataset, 28,105 reads for the soil dataset, and 12,481 reads for the animal dataset). Unless otherwise noted, all analyses were performed on chimera-checked data. To explore the impact of read length on the detection of microbial alpha diversity, the 5 control samples of each dataset were selected to measure richness and inverse Simpson diversity (Chao, Chiu, & Jost, 2014), which are more heavily weighted by the rare and dominant taxa, respectively. Similarly, to explore the effects of read length on beta diversity, Bray-Curtis and Sorensen dissimilarities between samples were examined. To assess the extent to which read length affected the ecological conclusions derived from the data, samples from before and (1 day) after disturbance for each dataset were compared. For alpha diversity, control and disturbed samples were compared using a Wilcoxon test, and for beta diversity, control and disturbed samples were compared using a PERMANOVA (adonis2) for each read length. Finally, to examine the loss of ecological information with read length, a mantel test of the dissimilarities (Bray-Curtis and Sorensen) between the longest read length (200 bp) and all shorter reads was performed for each dataset.

Results

The sequence data recovered from INSDC databases ranged in average read length, between 200 bp (animal microbiome) to 300 bp (aquatic microbiome) of average read lengths and pre-processing steps (Figure 1). Notably, the aquatic microbiome data was archived with merged paired ends, for a total average length of 600 bp. On average across datasets, read qualities remained high until 200 bp, but were lowest for the soil microbiome data (Figure 1). Among the 15 samples included for each dataset, the number of reads per sample varied between 14,088 and 97,218 reads per sample animal, between 73,058 and 31,237 reads per sample for aquatic, and between 196,518 and 51,092 reads per sample for soil microbiomes.

Processing sequence data across a gradient of read lengths had consistent effects across datasets, with the average number of reads conserved per sample decreasing gradually with read length in data processed without chimera checking (Figure 2, top), consistent with the quality profile (Figure 1). Importantly, for lower quality sequences, trimming reads below 100 bp resulted in the removal of a large proportion of the original reads during chimera checking, but this was dependent on sequence quality (Figure 2, bottom). In the soil data, chimera checking removed over 25% of the original reads when trimming resulted in reads of 50-80 bp in length, but removed only 11.9 $\pm$ 8% if the reads were trimmed to 100 bp. In the higher quality animal microbiome dataset, chimera checking reads that were trimmed below 60 bp in length resulted in a

loss of 10.9±2%, while chimera checking reads that were 70 bp or longer only resulted in a loss of 5.1% of the original reads.

The percentage of taxonomically unclassified reads decreased sharply with read length. As expected, this effect was most pronounced at the lower taxonomic levels, but depended on the diversity and/or prior characterization of the system (Figure 3). For the less diverse and more well characterized animal microbiome, read lengths beyond 70 bp achieved minimal improvements in taxonomic classification, while aquatic and soil microbiomes achieved minimal improvements in taxonomic classification beyond 90 and 110 bp respectively. Genus-level classification followed a similar pattern, but required longer reads to reach saturation: in the animal microbiome, 87.1% of the community was classified on average when reads were 120 bp or longer, while in aquatic and soil microbiomes, trimming to 200 bp achieved a classification of 78.8% and 69.3% of the community, respectively.

The number of ASVs detected across the 5 control samples in each dataset was assessed given the same read depth and increasing read length (Figure S1). With increasing read depth, all datasets detected higher numbers of ASVs. Trimming reads to 50 bp resulted in the detection of 180 animal, 228 aquatic, and 367 soil ASVs, which increased to 1029, 1833, and 5105 ASVs, respectively, when trimming to 200 bp.

To determine how trimming affected the detection of diversity, the richness and inverse Simpson index in the 5 control samples of each dataset were assessed. Within each dataset, alpha diversity increased with read depth, but saturated more rapidly in less diverse environments (Figure 4). Richness estimates exhibited a hump with increasing read length in soil samples, decreasing after 160 reads in line with the decrease in read quality (Figure 1). Importantly, while similar patterns were observed for both alpha diversity estimates, Inverse Simpson's index was more robust to read lengths, saturating with 70, 100, and 160 bp in the animal, water, and soil samples.

The extent to which shorter read lengths affected the variance in compositional metrics (i.e., beta diversity) was assessed by measuring the pairwise Sorensen and Bray-Curtis dissimilarities between the five control samples in each dataset. For all datasets, increasing read lengths resulted in gradual, but saturating increases in dissimilarity. The point of saturation depended on the expected diversity in each system: animal and soil microbiomes approached saturation with 60 and 90 bp reads, respectively (Figure 5). Bray-Curtis dissimilarities were less variable (standard deviation of 0.03, 0.04, and 0.03 for animal, aquatic, and soil microbiomes, respectively) across read lengths than Sorensen dissimilarities (standard deviation of 0.03, 0.09, and 0.05 for animal, aquatic, and soil microbiomes, respectively). To further examine information loss from shorter read lengths, Mantel tests between the communities resulting from each read length and the dataset trimmed to 200 bp (i.e., the most information-rich version, Figure 6) were performed using Sorensen and Bray-Curtis dissimilarities. While the strength of Sorensen-based correlations increased with read length, Bray-Curtis dissimilarities were more robust, and exhibited little deviation from the 200 bp dataset. Importantly, this pattern was consistent regardless of whether a chimera-checking step was included (Figure S2).

Finally, the extent to which shorter read lengths recovered differences in alpha and beta diversity between the control and disturbed samples (1 day after disturbance) was evaluated for each dataset (Figure 7) using Wilcoxon rank sum tests and PERMANOVAs. In general, longer read lengths were able to better discriminate between the alpha diversity of control and disturbed samples, both in terms of richness and Inverse Simpson diversity (Figure 7, top). Above 100 bp, only marginal differences in discrimination between the richness in the control and disturbed treatments were detected with increasing read length for both metrics, with p-values ranging between 0.095-0.055, 0.012-0.010, and 0.151-0.309 for aquatic, animal, and soil microbiomes. respectively. Inverse Simpson diversity was more robust to decreasing lengths, and discrimination did not change with reads greater than 90 bp ($p=0.008$ all the comparison between control and disturbed in all microbiomes). The discrimination of beta diversity between control and disturbed samples exhibited similar patterns, with the exception of the water dataset evaluated with Sorensen dissimilarities, which indicated that samples became more similar with longer reads (Figure 7, bottom). Sorensen dissimilarity between the treatments in the aquatic communities decreased linearly with read length and discrimination decreased, from $p=0.074$ with 50 bp data to $p=0.456$ with 200 bp data). The abundance-weighted Bray-Curtis dissimilarities

were more robust to read lengths, exhibiting low variation in general.

Discussion

Amplicon sequencing remains the most common method for identifying microbial communities, largely due to its low price and high throughput relative to more novel techniques (e.g., long-read sequencing, shotgun metagenomics). As the popularity of amplicon sequencing continues to grow, so does the wealth of archived 16S rRNA sequences, and understanding how bioinformatics choices affect the definition of species, and how this in turn affects the detection of microbial diversity and changes in this diversity is essential for the interpretation and reuse of these data (Jurburg et al., 2022). This work evaluated how shorter read lengths affect the detection of microbial taxa, their taxonomic assignments, and biodiversity estimates derived from these data. Its findings indicate that short read lengths recover biodiversity patterns, but special caution should be taken in the selection of biodiversity metrics to examine these data.

As expected, shorter read lengths resulted in more unclassified ASVs, but this was dependent on the target taxonomic level and varied across read lengths. Classification was best in the animal dataset, which was the least diverse and most-well characterized system. Importantly, only marginal improvements in taxonomic assignments were obtained by read lengths greater than 100 bp at the family level and above for all the datasets used, suggesting that, if only forward reads are available, little information is lost by 100 bp reads relative to the full forward read. Our results also highlight that genus-level taxonomic assignments greatly depend on how well-characterized the microbiota of the target environment are, and suggest that interpretations of genus-level assignments are not recommended for shorter reads (Thompson et al., 2017).

Further analyses highlighted the robustness of alpha and beta diversity metrics, especially abundance-weighted metrics (i.e., Inverse-Simpson index and Bray-Curtis dissimilarities, to shorter read lengths. Reads of 90 bp could recover the majority of the alpha diversity observed with 200 bp, as well as the dissimilarity between communities belonging to both biological replicates (i.e., variance or dispersion) and different treatments. Importantly, the similarity between the 200 bp datasets and their shorter versions increased with read length when assessed with incidence-based Sorensen dissimilarities, but remained high for abundance-weighted Bray-Curtis dissimilarities, even for the shortest reads. As these two dissimilarity metrics differ only in their abundance weighing, the differences observed when using each suggest that rare taxa are the ones most affected by shorter read lengths, highlighting the dependence of rare taxa on bioinformatics parameters.

Similarly, the detection of ASVs increased linearly with read length until a saturation point that aligned with the expected diversity in each environment explored (i.e., from least to most diverse, the animal, aquatic, and soil microbiomes), emphasizing the importance of defining diversity estimates relative to the trimming parameters. These results highlight the importance of considering diversity estimates, particularly incidence-based alpha diversity metrics (i.e., richness) as a function of read length. In the case of data reuse and comparison among datasets, this study demonstrates the importance of applying a uniform read length across datasets in order to have comparable diversity estimates.

With second generation sequence data (i.e., Illumina MiSeq), sequence quality decreases with read length (Ben J Callahan et al., 2016). Consequently, less reads pass quality checking, resulting in less reads (or observations) in the final, processed dataset. Short read lengths may therefore increase the number of observations per sample, particularly in low-quality sequences. Furthermore, different studies employ different sequencing platforms, which produce reads of variable lengths, the shortest of which is Illumina HiSeq, featuring a maximum read length of 150 bp, including barcodes and primers (Di Bella, Bao, Gloor, Burton, & Reid, 2013). In the case of pair-ended sequence data, only forward or merged reads are often archived (Jurburg et al., 2020). This work demonstrates how one aspect of sequence processing (i.e., trimming) affects the detection and taxonomic assignment of microbial diversity. While several studies have examined how technical choices (i.e., primer choice (Fouhy, Clooney, Stanton, Claesson, & Cotter, 2016; Martínez-Porchas, Villalpando-Canchola, & Vargas-Albores, 2016; Tremblay et al., 2015), pipeline selection (Marizzoni et al., 2020), and rarefaction (McKnight et al., 2018; Weiss et al., 2017)) affect the detection of diversity, systematic assessments of how other technical choices (particularly bioinformatics parameters e.g., chimera checking)

affect the microbial diversity estimates are lacking, but urgently needed. Importantly, short reads enable the reuse of sequence data in their rawest form, allowing for complete and unified reprocessing of the sequence data from different studies, which may in turn improve comparability among them (Kang et al., 2021).

Processing metabarcoding data requires making a series of choices that affect the final dataset and its interpretation (Abellan-Schneyder et al., 2021). Sequence trimming is a critical part of processing, but its effect on the resulting diversity estimates are often overlooked. The analyses presented focused on the effect of sequence trimming in the popular *dada2* pipeline, which detects amplicon sequence variants (ASVs) rather than grouping sequences into clusters of 97% sequence similarity. *Dada2* has been extensively validated, and exhibits high sensitivity to ASVs (Prodan et al., 2020). While the findings in this study may guide the general processing of amplicon sequencing data, it is important to note that the findings are specific to the *dada2* pipeline.

This study lays the groundwork for the analysis and reanalysis of metabarcoding data using short read lengths, and results in several recommendations. First, when comparing data with different technical backgrounds (i.e., from different studies), trimming to the same read length is important, especially for the analysis of alpha diversity. Second, when using short read lengths, caution should be taken with the interpretation of genus-level classifications. Third, abundance-weighted diversity metrics (i.e., inverse Simpson index, Bray-Curtis dissimilarity) are more robust to read length than incidence-based metrics (i.e., richness and Sorensen dissimilarity). Finally, the detection of microbial diversity from sequence data is far from absolute, and should instead be considered relative to the read length employed.

Acknowledgements

We would like to thank A. Chatzinotas and S. Tem for the valuable discussions. We acknowledge support by the German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118, 202548816). This study has been partly performed using the High-Performance Computing (HPC) Cluster EVE, a joint effort of both the Helmholtz Centre for Environmental Research - UFZ and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig.

Bibliography

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., ... Neuhaus, K. (2021). Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *MSphere*, *6*(1). doi: 10.1128/mSphere.01202-20

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., ... Buck, G. A. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, *15*, 66. doi: 10.1186/s12866-015-0351-6

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*,*6*, 190007. doi: 10.1038/sdata.2019.7

Callahan, Benjamin J, McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, Ben J, Sankaran, K., Fukuyama, J. A., McMurdie, P. J., & Holmes, S. P. (2016). Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. [version 2; peer review: 3 approved]. *F1000Research*, *5*, 1492. doi: 10.12688/f1000research.8986.2

Callahan, B J, McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. doi: 10.1038/nmeth.3869

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., ...

Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.*Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 1*(Suppl 1), 4516–4522. doi: 10.1073/pnas.1000080107

Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 297–324. doi: 10.1146/annurev-ecolsys-120213-091540

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.*Clinical Microbiology Reviews*, *17*(4), 840–862, table of contents. doi: 10.1128/CMR.17.4.840-862.2004

Curry, K. D., Wang, Q., Nute, M. G., Tyshaieva, A., Reeves, E., Soriano, S., ... Treangen, T. J. (2022). Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nature Methods*, *19*(7), 845–853. doi: 10.1038/s41592-022-01520-4

Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P., & Reid, G. (2013). High throughput sequencing methods and analysis for microbiome research.*Journal of Microbiological Methods*, *95*(3), 401–414. doi: 10.1016/j.mimet.2013.08.011

Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J., & Cotter, P. D. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology*, *16*(1), 123. doi: 10.1186/s12866-016-0738-z

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., ... Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. doi: 10.1038/s41467-019-13036-1

Jurburg, S. D., Buscot, F., Chatzinotas, A., Chaudhari, N. M., Clark, A. T., Garbowski, M., ... Heintz-Buschart, A. (2022). The community ecology perspective of omics data. *Microbiome*, *10*(1), 225. doi: 10.1186/s40168-022-01423-8

Jurburg, S. D., Cornelissen, J. J. B. W. J., de Boer, P., Smits, M. A., & Rebel, J. M. J. (2019). Successional Dynamics in the Gut Microbiome Determine the Success of Clostridium difficile Infection in Adult Pig Models.*Frontiers in Cellular and Infection Microbiology*, *9*, 271. doi: 10.3389/fcimb.2019.00271

Jurburg, S. D., Konzack, M., Eisenhauer, N., & Heintz-Buschart, A. (2020). The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Communications Biology*, *3*(1), 474. doi: 10.1038/s42003-020-01204-9

Jurburg, S. D., Nunes, I., Stegen, J. C., Le Roux, X., Priemé, A., Sørensen, S. J., & Salles, J. F. (2017). Autogenic succession and deterministic recovery following disturbance in soil bacterial communities. *Scientific Reports*, *7*, 45691. doi: 10.1038/srep45691

Kang, X., Deng, D. M., Crielaard, W., & Brandt, B. W. (2021). Reprocessing 16S rRNA Gene Amplicon Sequencing Studies: (Meta)Data Issues, Robustness, and Reproducibility. *Frontiers in Cellular and Infection Microbiology*, *11*, 720637. doi: 10.3389/fcimb.2021.720637

Louca, S., Mazel, F., Doebeli, M., & Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biology*,*17*(2), e3000106. doi: 10.1371/journal.pbio.3000106

Marizzoni, M., Gurry, T., Provasi, S., Greub, G., Lopizzo, N., Ribaldi, F., ... Cattaneo, A. (2020). Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples. *Frontiers in Microbiology*, *11*, 1262. doi: 10.3389/fmicb.2020.01262

Martínez-Porchas, M., Villalpando-Canchola, E., & Vargas-Albores, F. (2016). Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon*,*2*(9), e00170. doi: 10.1016/j.heliyon.2016.e00170

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., ... Hirota, K. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION^{TM} nanopore sequencing confers species-level resolution. *BMC Microbiology*,*21*(1), 35. doi: 10.1186/s12866-021-02094-5

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2018). Methods for normalizing microbiome data: an ecological perspective. *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.13115

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *Plos One*,*8*(4), e61217. doi: 10.1371/journal.pone.0061217

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., ... Wagner, H. (2007). The vegan package. *Community Ecology*.

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *Plos One*, *15*(1), e0227434. doi: 10.1371/journal.pone.0227434

Qian, J., Ding, Q., Guo, A., Zhang, D., & Wang, K. (2017). Alteration in successional trajectories of bacterioplankton communities in response to co-exposure of cadmium and phenanthrene in coastal water microcosms.*Environmental Pollution*, *221*, 480–490. doi: 10.1016/j.envpol.2016.12.020

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*,*41*(Database issue), D590-6. doi: 10.1093/nar/gks1219

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.*Plos One*, *6*(12), e27310. doi: 10.1371/journal.pone.0027310

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., ... Earth Microbiome Project Consortium. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*,*551*(7681), 457–463. doi: 10.1038/nature24621

Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., ... Tringe, S. G. (2015). Primer and platform effects on 16S rRNA tag sequencing.*Frontiers in Microbiology*, *6*, 771. doi: 10.3389/fmicb.2015.00771

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., ... Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. doi: 10.1186/s40168-017-0237-y

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090. doi: 10.1073/pnas.74.11.5088

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, *51*(2), 221–271. doi: 10.1128/mr.51.2.221-271.1987

Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*, *17*, 135. doi: 10.1186/s12859-016-0992-y

Yu, Z., García-González, R., Schanbacher, F. L., & Morrison, M. (2008). Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by Archaea-specific PCR and denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology*, *74*(3), 889–893. doi: 10.1128/AEM.00684-07

Zhang, W., Fan, X., Shi, H., Li, J., Zhang, M., Zhao, J., & Su, X. (2023). Comprehensive Assessment of 16S rRNA Gene Amplicon Sequencing for Microbiome Profiling across Multiple Habitats. *Microbiology Spectrum*, *11*(3), e0056323. doi: 10.1128/spectrum.00563-23

Figure captions

Figure 1. Quality profiles for selected samples in each study. Each panel shows the quality profile for the sample with the highest (top) and lowest (bottom) number of reads for each study. The gray scale indicates the frequency of each quality score at each base position (darkness indicates a higher frequency). Green and orange lines indicate the mean and quartile quality score at each position, respectively.

Figure 2. Percentage of reads preserved after standard processing with dada2 (Ben J Callahan et al., 2016) without (top) and with (bottom) chimera checking. Control samples (n=5) are shown for each study, as a percent of the original reads recovered from INSDC databases.

Figure 3. Relationship between taxonomic classification and read length, for the control samples (n=5) of each dataset. Color indicates taxonomic level.

Figure 4. Relationship between alpha diversity and read length. Richness (q=0, top) was calculated as the number of ASVs per sample. Inverse Simpson's index (q=2, bottom) was calculated according to (Chao et al., 2014). The diversity in control samples (n=5) was assessed for each read length.

Figure 5. Variance in community composition with increasing read lengths. The mean pairwise dissimilarity between the 5 control samples in each study was assessed using Sorensen (a) and Bray-Curtis (b) dissimilarities.

Figure 6. Information loss from shorter read lengths. For each dataset, Mantel tests between the Sorensen (a) and Bray-Curtis (b) dissimilarities 200-bp reads and each shorter read length evaluated the correlation in microbial communities between shorter read lengths and the most information-rich version of the dataset (200 bpp).

Figure 7. Outcome of statistical tests comparing control and disturbed communities for each dataset, across read lengths. Kruskall-Wallis tests (top) evaluated differences in richness and inverse Simpson's index between the control and recently-disturbed (1 day, n=5 for each time point) for each study. Similarly, PERMANOVA tests (bottom) evaluated differences in the community composition between control and recently-disturbed samples using Sorensen and Bray-Curtis dissimilarities.

Data accessibility and benefit sharing

Data for this study was downloaded from NCBI's Sequence Read Archives under accession numbers PRJNA329541, PRJDB4992, and PRJNA528235. Code and metadata used to generate the analyses presented is available in*https://github.com/drcarrot/ReadLength*.

Author contributions

SDJ designed the study, selected the data, performed the analyses, and wrote the manuscript.

Conflict of interest

The author declares no conflict of interest.