A Comprehensive Evaluation of Single-end Sequencing Data Analyses for Environmental Microbiome Research

Meganathan Ramakodi (✉ meganathan.pr@gmail.com)

National Environmental Engineering Research Institute CSIR <https://orcid.org/0000-0002-0034-6306>

Research Article

Keywords: Environmental microbiome, amplicon sequencing, Illumina MiSeq platform, single-end sequencing data, low quality reads, taxonomy inference

Posted Date: June 10th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-593687/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Archives of Microbiology on October 15th, 2021. See the published version at <https://doi.org/10.1007/s00203-021-02597-9>.

Abstract

Illumina sequencing platforms have been widely used for amplicon-based environmental microbiome research. Analyses of amplicon data of environmental samples, generated from Illumina MiSeq platform illustrate the reverse (R2) reads in the PE datasets to have low quality towards the 3' end of the reads which affect the sequencing depth of samples and ultimately impact the sample size which may possibly lead to an altered outcome. This study evaluates the usefulness of single-end (SE) sequencing data in microbiome research when the Illumina MiSeq PE dataset shows significantly high number of low quality reverse reads. In this study, the amplicon data (V1V3, V3V4, V4V5 and V6V8) from 128 environmental (soil) samples, downloaded from SRA, demonstrate the efficiency of single-end (SE) sequencing data analyses in microbiome research. The SE datasets were found to infer the core microbiome structure as comparable to the PE dataset. Conspicuously, the forward (R1) datasets inferred a higher number of taxa as compared to PE datasets for most of the amplicon regions, except V3V4. Thus, analyses of SE sequencing data, especially R1 reads, in environmental microbiome studies could ameliorate the problems arising on sample size of the study due to low quality reverse reads in the dataset. However, care must be taken while interpreting the microbiome structure as few taxa observed in the PE datasets were absent in the SE datasets. In conclusion, this study demonstrates the availability of choices in analyzing the amplicon data without having the need to remove samples with low quality reverse reads.

Introduction

Amplicon-based microbiome approach has been widely employed to understand the community structure and role of microorganisms in environmental research. In amplicon microbiome studies, different hypervariable regions of 16S rRNA gene such as V1V3, V3V4, V4V5 or V6V8 are generally amplified and sequenced on high-throughput sequencing (HTS) platforms. Today, researchers have a wide-variety of HTS platforms among which Illumina MiSeq has been adopted as a popular tool to generate amplicon data (Pollock et al. 2018; Singer et al. 2019). The Illumina MiSeq has also been used in the large scale environmental studies such as the Earth Microbiome Project, global soil analyses, climate change research, etc (Gilbert et al. 2014; Thompson et al. 2017; Bižić et al. 2020; Oliverio et al. 2020). Although the Illumina MiSeq platform has advantages such as low cost, flexibility, fast run time and generation of 300 bp long paired-end (PE) sequencing reads (Wen et al. 2017; Bharti and Grimm 2021), the platform has some disadvantages. One such disadvantage is having low phred quality towards the 3' end of the reads (Liu et al. 2020). Especially, the quality scores of R2 reads (reverse) are often lower than that of R1 reads (forward) (Schirmer et al. 2015; Tan et al. 2019). This study also evaluated the quality scores of PE sequencing data of 536 environmental samples, sequenced on Illumina MiSeq platform; available on SRA, and the results are shown in Fig. 1. The analyses revealed that the phred quality score of approximately 30–40% of R2 reads is reduced below 20 after ~ 225 bp.

In a standard bioinformatics workflow, the sequencing reads are primarily trimmed at the 3' end to remove the poor quality bases (Ex: phred quality < 20) so as to retain maximum number of reads for downstream analyses. Later, the R1 and R2 reads are stitched together to obtain the complete amplicon.

However, a copious amount of R2 reads, generated by MiSeq platform, often have low phred score after ~ 225 bp as observed in Fig. 1 and trimming of R2 reads at 225 bp could lead to a significant reduction of overlapping region which will affect the merging process of R1 and R2 reads. Another option is trimming the reads after spanning enough overlapping region for merging but still a sizeable population of R2 reads will be removed from the dataset when stringent quality criteria is used during the filtering step. Thus, the low quality of R2 reads in PE sequencing data of samples could lead to significant reduction in sequencing depth of respective samples during the filtering and merging processes. Earlier studies have highlighted that the microbiome composition and structure could be influenced by sequencing depth (Zaheer et al. 2018; on behalf of the REHAB consortium et al. 2019; Ramakodi 2021). Thus, samples with low sequencing depth due to poor quality of R2 reads need to be removed from the dataset to avoid problematic results. In contrast, most of the R1 reads of samples have good quality (phred score > 20) bases even after 275 bp (Fig. 1). Thus, exclusively analyzing the R1 reads could help to retain most or all the samples in microbiome analyses. Earlier, Werner et al. (2012) compared the Illumina PE and SE reads using 2X150 bp PE data of V4 region and their study illustrated that the alpha and beta diversity indices based on SE sequencing data are comparable to the results derived from PE data. However, the efficiency of SE data for taxonomy profiling was not evaluated. In addition, Werner et al. (2012) analyzed 2X150 bp data of different samples that was available at that time but today the Illumina MiSeq platform has the chemistry which generates up to 300 bp reads. However, the efficiency of single-end (SE) sequencing read (R1) analyses to infer microbiome structure in environmental research is not yet known. Thus, a comprehensive evaluation of SE and PE reads generated from Illumina MiSeq platform to assess microbiome structure is required at this moment. In this context, this study aims to evaluate the effectiveness of SE sequencing read analyses in environmental microbiome research.

Materials And Methods

Data

The V1V3, V3V4, V4V5 and V6V8 amplicon sequencing data, generated by Soriano-Lerma et al. (2020) for soil samples, were analyzed in this study to evaluate the usefulness of SE sequencing reads for environmental microbiome research. The location of sampling, sample collection protocol and amplicon sequencing procedure are described in detail in Soriano-Lerma et al. (2020). Briefly, a total number of 128 samples (32 samples for each amplicon region), as submitted by the authors were downloaded from the Sequence Reads Archive (SRA) of NCBI for the analyses (BioProject Accession Number: PRJNA612815). All the data were downloaded using the SRA toolkit. The SRA accession of each sample along with their metadata as obtained from SRA are given in Supplementary Table 1. The samples were divided into four groups according to the 16S rRNA amplicon region and the dataset were analyzed, separately.

Amplicon microbiome analyses

The amplicon sequencing data was analyzed on R version 3.6.3 using DADA2 tool which generates Amplicon Sequencing Variants (ASVs). Recent studies have highlighted that ASVs could infer the

microbiome structure better than the conventional Operational Taxonomic Units (OTUs) (Callahan et al. 2016; Caruso et al. 2019) which is based on clustering approach. Thus, this study adopted ASVs to analyze the environmental microbiome data. The schematic diagram showing the DADA2 workflow along with the parameters used are shown in Supplementary Fig. 1. Briefly, the R1, R2 and PE (R1 and R2) data of samples were analyzed, separately. Subsequently, the non-chimeric sequences were generated based on (i) only R1 reads, (ii) only R2 reads and (iii) PE data. Thus, this study generated three ASV datasets for each amplicon sequencing region for further downstream analyses. The data generated from R1 or R2 are referred as SE dataset whereas the data generated by merging R1 and R2 reads are referred as PE dataset.

Taxonomy assignment

The taxonomy assignments of ASVs were carried out using the tool IdTaxa (Murali et al. 2018). Briefly, each dataset was analyzed, separately, and a stringent parameter 70% confidence threshold was selected for the inference of taxonomy. The SILVA SSU database 138 (www.arb-silva.de) as available on the IdTaxa web tool was used as the training dataset for the taxonomic assignment of ASVs.

Downstream analyses

The downstream analyses were performed in R version 3.6.3 using the packages phyloseq (McMurdie and Holmes 2013), microbiome, Decipher (Wright 2016), ggplot2 (Wickham 2016), tidyverse (Wickham et al. 2019), dplyr, ape (Paradis et al. 2004), vegan, hclust and venn. Primarily, the ASVs without known phylum were removed from the dataset. Also, the ASVs were removed, if the ASVs exist in only one sample and/or have less than 0.001 proportion of minimum sample depth in the dataset. The alpha and beta diversity analyses were performed using the rarefied dataset. The minimum sequencing depth of the dataset was used as the rarefaction depth. The core microbiome structure was inferred based on the following criteria that the taxa have more than 0.001 of relative abundance with the prevalence of 50%.

Results And Discussion

The SE sequencing read analyses in microbiome research was evaluated using the data generated from the Illumina MiSeq platform which is widely used for amplicon microbiome research worldwide. The results of this current study support the usefulness of SE data analyses in environmental microbiome studies.

Quality of reverse reads influences the sample sequencing depth

Primarily, the impact of R2 read qualities on generation of full length amplicon was analyzed. Briefly, the R1 and the R2 reads were truncated at different lengths, the PE data was merged and the results were compared (Fig. 2). The results show that the depth (number of sequences/ reads of a sample) of R2 and merged sequences is inversely proportional to the truncation length of R2 reads whereas the truncation

length of R1 reads does have a great impact on merging the PE dataset. The results indicate that the quality of R2 reads, obtained from Illumina MiSeq platform, decreases as the length of reads increases and the low base quality of R2 reads affects the sequencing depth of samples, irrespective of amplicon regions. A similar observation of decrease in the quality of R2 reads as compared to R1 reads were reported in the earlier studies (Chen et al. 2018; Liu et al. 2020). Thus, low quality of R2 reads in Illumina MiSeq dataset is a consistent phenomenon which leads to the reduction in sequencing depth of samples which in turn could affect the environmental microbiome studies (Pereira-Marques et al. 2019; Ramakodi 2021).

SE versus PE datasets for the inference of alpha and beta diversity

This study evaluated the R1, R2 and PE datasets of four different amplicons, V1V3, V3V4, V5V5, and V6V8, for inferring the microbiome structure in environmental microbiome research. The analyses showed that the number of non-chimeric sequences obtained for R1 datasets (Median: V1V3- 43,333; V3V4- 33,063; V4V5- 37,313; V6V8- 34,932) is higher than that of R2 (Median: V1V3- 19,166; V3V4- 30,892; V4V5- 20,390; V6V8- 21,825) and PE (Median: V1V3- 9,579; V3V4- 16,539; V4V5- 20,257; V6V8- 19,652) datasets. Similarly, the observed ASVs were found to be significantly higher (P-value: $< 2.2e-16$ to 0.012) for R1 datasets as compared to the other datasets, irrespective of amplicon regions. The comparison of distribution of observed ASVs based on R1, R2 and PE datasets are shown in Fig. 3. Earlier studies showed that the sequencing depth could influence the number of ASVs (Ramakodi 2021). Thus, the higher distribution of observed ASVs in R1 datasets as compared to its counterparts could be due to the higher sequencing depth in R1 datasets. The beta diversity of samples based on bray-curtis dissimilarity index was studied and the dendrogram based on the bray-curtis distance values are shown in Supplementary Figs. 2 to 5 for V1V3, V3V4, V5V5, and V6V8, respectively. The beta diversity analyses show that the relationships between samples based on R1, R2 and PE datasets vary, irrespective of amplicon regions. The bray-curtis distance is based on the microbial community composition of datasets. The alpha diversity results clearly illustrate R1, R2 and PE datasets to exhibit different proportions of ASVs within each amplicon region. Thus, the differences in the distribution of the number of ASVs/ microbial taxa within R1, R2 and PE datasets could have influenced beta diversity.

Core microbiome composition: SE versus PE datasets

The ultimate goal of any microbiome study is to find the structure of core microbiome and this study also evaluated the utility of SE sequencing data to infer the core microbiome composition. The comparison of core microbiome at genus level, based on SE and PE datasets of different amplicon regions is shown in Fig. 4 and the relative abundance of taxa (class level) is shown in Fig. 5. The analyses showed the R1 datasets to have a higher number of unique genus as compared to R2 and PE datasets. Similarly, the R1 datasets exhibited higher numbers of unique class, order and family level taxa for all the amplicon regions, except V3V4 (Supplementary Fig. 6). The results suggest that the SE datasets, especially the datasets comprising exclusively R1 reads, could infer more number of taxa as compared to PE datasets. The reason for observing a higher number of taxa in R1 datasets could be attributed to the sequencing

depth. Earlier studies highlighted that the microbiome data is compositional in nature which means the observed microbiome structure is defined by the number of reads available in the dataset (Gloor et al. 2017; Susin et al. 2020). A higher sequencing depth could infer more taxa as compared to the dataset having low sequencing depth (Zaheer et al. 2018). In this study, the R1 datasets had a higher number of non-chimeric sequences. Thus, the R1 datasets yielded a higher number of taxa. These observations suggest that the R1 reads could provide more information on the microbiome including the rare taxa which require higher sequencing depth. However, some discrepancies in inferring the core microbiome were observed between SE and PE sequencing data which could be attributed to the shorter length of SE sequencing data which affects the phylogenetic resolution (Fuks et al. 2018; Johnson et al. 2019). Also, the microbiome composition observed for different amplicon regions vary which is not surprising as the original study from which the data was adopted herein, also illustrated the discrepancies on microbiome composition by different amplicon regions (Soriano-Lerma et al. 2020). In summary, this study suggests that the SE sequencing data, especially the R1 reads, provide results comparable to PE datasets. In fact, the R1 datasets yielded more taxa as compared the PE datasets, irrespective of amplicon regions. Thus, the SE sequencing data analyses could be adopted as an alternative approach to infer microbiome composition as and when the quality of R2 reads, generated by Illumina MiSeq platform, are low and significantly reduces the PE data and subsequently, affects the sample size of the study. However, the results obtained from SE sequencing datasets need careful interpretation as some of the taxa observed in PE datasets were absent in SE datasets.

Conclusions

The Illumina MiSeq platform often yields low quality R2 reads which affects the sample sequencing depth. In general, the samples with low sequencing depth need to be removed from the dataset to control the bias due to sequencing depth. Nonetheless, removing the samples could lead to spurious results and also may affect the study design. In this context, the researchers need to adopt an alternative approach to infer the microbiome structure to overcome the problems associated with low quality data. This study demonstrates that the SE sequencing datasets, especially the R1 reads, infers the core microbiome structure as comparable to the PE dataset. Conspicuously, the R1 datasets inferred a higher number of taxa as compared to PE datasets, irrespective of amplicon regions. Thus, analyses of SE sequencing data, especially R1 reads, in amplicon microbiome studies could ameliorate the problems related to low quality reverse reads in the dataset. However, care must be taken while interpreting the microbiome structure as the SE sequencing data could not resolve the taxonomy of some of the taxa. In conclusion, this study demonstrates the SE data analysis as one of the workflow that could be adopted in amplicon microbiome analyses to overcome sample elimination issues arising due to problematic reverse reads.

Declarations

ACKNOWLEDGEMENTS:

I would like to thank Dr. Bhawna Dubey, Chief Scientific Officer, Reprocell Bioserve Biotechnologies Pvt. Ltd., Hyderabad for reviewing the manuscript. The effort of Soriano-Lerma et al. (2020) for making the data publicly available on SRA is highly appreciated. CSIR-NEERI is acknowledged for providing the necessary support to carry out the analyses. The manuscript draft is submitted in the Institute Repository under the KRC No.: CSIR-NEERI/KRC/2021/MAY/HZC/3

Funding:

Not applicable

Conflicts of interest/Competing interests:

Not applicable

Availability of data and material:

Downloaded from SRA

Code availability:

Not applicable

Authors' contributions:

Single author

Ethics approval:

Not applicable

Consent to participate:

Not applicable

Consent for publication:

Not applicable

References

1. Bharti R, Grimm DG (2021) Current challenges and best-practice protocols for microbiome analysis. *Brief Bioinform* 22:178–193. <https://doi.org/10.1093/bib/bbz155>
2. Bižić M, Klintzsch T, Ionescu D, et al (2020) Aquatic and terrestrial cyanobacteria produce methane. *Sci Adv* 6:eaax5343. <https://doi.org/10.1126/sciadv.aax5343>

3. Callahan BJ, McMurdie PJ, Rosen MJ, et al (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>
4. Caruso V, Song X, Asquith M, Karstens L (2019) Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems* 4:e00163-18, /msystems/4/1/msys.163-18.atom. <https://doi.org/10.1128/mSystems.00163-18>
5. Chen X, Johnson S, Jeraldo P, et al (2018) Hybrid-denovo: a de novo OTU-picking pipeline integrating single-end and paired-end 16S sequence tags. *Gigascience* 7:1–7. <https://doi.org/10.1093/gigascience/gix129>
6. Fuks G, Elgart M, Amir A, et al (2018) Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 6:17. <https://doi.org/10.1186/s40168-017-0396-x>
7. Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. *BMC Biol* 12:69. <https://doi.org/10.1186/s12915-014-0069-1>
8. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* 8:2224. <https://doi.org/10.3389/fmicb.2017.02224>
9. Johnson JS, Spakowicz DJ, Hong B-Y, et al (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 10:5029. <https://doi.org/10.1038/s41467-019-13036-1>
10. Liu T, Chen C-Y, Chen-Deng A, et al (2020) Joining Illumina paired-end reads for classifying phylogenetic marker sequences. *BMC Bioinformatics* 21:105. <https://doi.org/10.1186/s12859-020-3445-6>
11. McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* 8:e61217. <https://doi.org/10.1371/journal.pone.0061217>
12. Murali A, Bhargava A, Wright ES (2018) IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 6:140. <https://doi.org/10.1186/s40168-018-0521-5>
13. Oliverio AM, Geisen S, Delgado-Baquerizo M, et al (2020) The global-scale distributions of soil protists and their contributions to belowground systems. *Sci Adv* 6:eaax8787. <https://doi.org/10.1126/sciadv.aax8787>
14. on behalf of the REHAB consortium, Gweon HS, Shaw LP, et al (2019) The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiome* 14:7. <https://doi.org/10.1186/s40793-019-0347-1>
15. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290. <https://doi.org/10.1093/bioinformatics/btg412>
16. Pereira-Marques J, Hout A, Ferreira RM, et al (2019) Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front*

- Microbiol 10:1277. <https://doi.org/10.3389/fmicb.2019.01277>
17. Pollock J, Glendinning L, Wisedchanwet T, Watson M (2018) The Madness of Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies. *Appl Environ Microbiol* 84:e02627-17, /aem/84/7/e02627-17.atom. <https://doi.org/10.1128/AEM.02627-17>
 18. Ramakodi MP (2021) Effect of Amplicon Sequencing Depth in Environmental Microbiome Research. *Curr Microbiol* 78:1026–1033. <https://doi.org/10.1007/s00284-021-02345-8>
 19. Schirmer M, Ijaz UZ, D’Amore R, et al (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* 43:e37–e37. <https://doi.org/10.1093/nar/gku1341>
 20. Singer GAC, Fahner NA, Barnes JG, et al (2019) Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci Rep* 9:5991. <https://doi.org/10.1038/s41598-019-42455-9>
 21. Soriano-Lerma A, Pérez-Carrasco V, Sánchez-Marañón M, et al (2020) Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Sci Rep* 10:13637. <https://doi.org/10.1038/s41598-020-70141-8>
 22. Susin A, Wang Y, Lê Cao K-A, Calle ML (2020) Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics* 2:lqaa029. <https://doi.org/10.1093/nargab/lqaa029>
 23. Tan G, Opitz L, Schlapbach R, Rehrauer H (2019) Long fragments achieve lower base quality in Illumina paired-end sequencing. *Sci Rep* 9:2856. <https://doi.org/10.1038/s41598-019-39076-7>
 24. Thompson LR, Sanders JG, McDonald D, et al (2017) A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* 551:457–463. <https://doi.org/10.1038/nature24621>
 25. Wen C, Wu L, Qin Y, et al (2017) Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS ONE* 12:e0176716. <https://doi.org/10.1371/journal.pone.0176716>
 26. Werner JJ, Zhou D, Caporaso JG, et al (2012) Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME J* 6:1273–1276. <https://doi.org/10.1038/ismej.2011.186>
 27. Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed. 2016. Springer International Publishing: Imprint: Springer, Cham
 28. Wickham H, Averick M, Bryan J, et al (2019) Welcome to the Tidyverse. *JOSS* 4:1686. <https://doi.org/10.21105/joss.01686>
 29. Wright E S (2016) Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal* 8:352. <https://doi.org/10.32614/RJ-2016-025>
 30. Zaheer R, Noyes N, Ortega Polo R, et al (2018) Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci Rep* 8:5890. <https://doi.org/10.1038/s41598-018-24280-8>

Figures

Figure 1

Distribution of base call quality scores of sequencing reads generated by Illumina MiSeq sequencing platform. (A) R1- Forward reads; (B) R2- Reverse reads

Figure 2

Effect of truncation length of sequencing reads on merging paired-end reads. The results suggest that the number of filtered R2 reads decreases as the truncation length of R2 reads increases

Figure 3

Distribution of observed ASVs based on SE and PE datasets for different amplicon regions.

Figure 4

Comparison of core microbiome taxa (genus level) between SE and PE datasets. The results suggest the R1 datasets infer a higher number of taxa for most of the amplicon region, except V3V4

Figure 5

Core microbiome structure (Class level) based on SE and PE datasets for different amplicon regions

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.pdf](#)