PLOS ONE

RESEARCH ARTICLE

Simple mapping-based quantification of a mock microbial community using total RNA-seq data

Shigeharu Moriya *

Environmental Metabolic Analysis Research Team, Center for Sustainable Resource Science, RIKEN Institute, Yokohama, Kanagawa, Japan

* smoriya@riken.jp

OPEN ACCESS

Citation: Moriya S (2021) Simple mapping-based quantification of a mock microbial community using total RNA-seq data. PLoS ONE 16(7): e0254556. https://doi.org/10.1371/journal.pone.0254556

Editor: Ruslan Kalendar, University of Helsinki: Helsingin Yliopisto, FINLAND

Received: July 21, 2020

Accepted: June 29, 2021

Published: July 16, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0254556

Data Availability Statement: All sequence data are deposited in DDBJ DRA, accession number DRA009985. (https://ddbj.nig.ac.jp/DRASearch/submission?acc=DRA009985).

Abstract

Most microbes in the natural environment are difficult to cultivate. Thus, culture-independent analysis for microbial community structure is important for the understanding of its ecological functions. An immense ribosomal RNA sequence collection is available from phylogenetic research on organisms in all domains. These sequences are available for use in genetic research. However, the amplicon-seq process using PCR requires the construction of a sequence library. Construction can introduce bias into quantitative analyses, and each domain of species needs its own primer set. Total RNA sequencing has the advantage of analyzing an entire microbial community, including bacteria, archea, and eukaryote, at once. Such analysis yields large amounts of ribosomal RNA sequences that can be used for analysis without PCR bias. Evaluation using total RNA-seq for quantitative analysis of microbial communities and comparison with amplicon-seq is still rare. In the present study, we developed a mapping-based total RNA-seq analysis to obtain quantitative information on microbial community structure and compared our results with ordinary amplicon-seq methods. We read total RNA sequences from a commercially available mock community (ATCC MSA-2003) and divided reads into small subunit ribosomal RNA (ssrRNA) origin reads and others, such as mRNA origin reads. We then mapped ssrRNA origin reads on annotated assembled contigs and obtained quantitative results under several analysis strategies. Removal of low complexity sequences, sorting ssrRNA with paired-in mode, and performing homology-based taxonomical assignments (BLAST+ or vsearch) showed superior outcomes to other strategies. Results with this approach showed a median relative abundance among ten mock community members of ~10%; ordinary amplicon-seq showed a much lower percentage. Thus, total RNA-seq can be a powerful tool for analyzing microbial community structure and is not limited to analyzing gene expression profiling of microbiomes.

Introduction

Understanding ecological services of microbial communities require knowledge of community composition. Most microbes are difficult to cultivate, yet microbial community structure

Funding: This work was supported by the Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Scientific Research on Innovative Areas, JP3308, and the RIKEN integrated symbiology program.

Competing interests: Author have no competitive interests.

analysis is an important tool for investigation of environmental microbial activity. As an alternative to cultivation, a molecular phylogenetic approach is widely used. RNA and DNA can be extracted from environmental samples without cultivation, and PCR with specific "bar-code" gene(s) can be used for phylogenetic classification of microbes. Accumulation of molecular phylogenetic information allows molecular classification based on "bar-code" gene sequence comparisons with molecular phylogeny.

Small subunit ribosomal RNA (ssrRNA) is a well-established "bar-code" gene for taxonomic identification because it is conserved among organisms with the same biological function. Early molecular phylogenetic work with ssrRNA sequences uncovered three life domains [1] and an unexpected variety of not-yet-cultivated microorganisms [2]. Long-term accumulation of ssrRNA sequences in the context of phylogenetic and taxonomic investigation led to the development of large public ribosomal RNA (rRNA) databases such as RDP, Greengenes, and Silva [3–5]. Using this common "bar-code" among all domain organisms, we can identify microbes by similarity with sequences stored in these databases, and we can classify new species using of phylogenetic analyses with those sequences.

Modern molecular biology and high throughput sequencing provide the opportunity to comprehensively evaluate microbial communities. Microbial rRNA sequences can be amplified from any environmental or clinical samples and can be read by a massively parallel sequencer [6–10]. rRNA databases can then be used to define microbial community structure by comparison among "bar-code" genes. This technique is called "amplicon-seq" and is widely used in microbial ecology [8–10].

Amplicon-seq is a powerful tool but has several weak points. rRNA has several conserved sequences, e.g., the stem region, yet universal primers across different domains are difficult to create. Hence, amplicon-seq should be applied separately among domains—bacteria, archaea, or eukaryotes. Furthermore, PCR introduces bias because of sequence differences among microbes. In some cases, contamination with environmental DNA is a problem. For example, frozen soil sample may include not only live microbes but also microbes destroyed by the freezing process. DNA from dead microbes can cause noise, for example, when investigating seasonal changes in arctic soil microbiomes. Total RNA-seq can be used to address such issues [11–13].

RNA-seq is also well-established for analyzing expressed genes. This "transcriptome analysis" is typically performed with mRNA enriched with complementary DNA (cDNA). rRNA is present in much higher amounts than mRNA [14]. However, current sequencing technology can distinguish mRNA, even in the presence of relatively large amounts of rRNA. Presently, no poly-A tail mRNA containing microbial community can be analyzed by total RNA-seq.

In the RNA-seq process, huge amounts of rRNA information are obtained. This information is used for taxonomic analysis. Arctic environmental microbiologists applied RNA-seq to solve DNA contamination issues [11, 12], and rumen microbiologists used the method to evaluate microbial communities composed of bacteria and ciliates [15, 16]. Therefore, several RNA-seq-based analysis methods are available.

Analysis pipeline work with ribo-tag [12, 15] typically uses reads as tag sequences to annotate and quantify bar codes for molecular classification [5]. Short-length reads are used for this analysis, and annotation resolution is limited, e.g., up to the level of order or family. Identifying up to the level of genus or species requires a low-throughput method such as clone library construction.

Conversely, mapping-based RNA sequences use a different principle to annotate and quantify reads [17–19]. In this case, reads are mapped onto reference sequences, such as ssrRNA database contents. Miss-mapping is still possible because of highly conserved sequences among organisms in the stem region, but finer annotation, e.g., genus level, is still possible.

[17–19]. Little study using mock communities is available to compare total RNA-seq and amplicon-seq approaches. [17, 20].

In the present study, a modified mapping-based all RNA information sequencing (ARI-seq) analysis using a mock microbial community was compared with an amplicon-seq analysis pipeline. We constructed contigs with the obtained reads and mapped these reads onto our own in-house total cDNA database. Simultaneously, we divided the reads into possible ssrRNA origin and others. We then expected that ssrRNA origin and "other RNA" (possibly mRNA and other functional RNA) reads are separately mapped in an in-house cDNA database. This simple process is slightly different from ordinary mapping-based RNA sequences in that reference sequences are constructed from their own reads instead of library contents. This approach is expected to add confidence and accuracy because reference sequences are directly generated from obtained reads.

Our results show that specific conditions of analysis are needed and that our method displays genus-level accuracy for taxonomic assignment. A mock community with ten species was correctly and quantitatively reproduced with assignments superior to amplicon-seq.

Materials and methods

Mock microbial community DNA and RNA preparation

We used ten strains of evenly mixed cell material (ATCC MSA-2003, American Type Culture Collection). The material includes well-characterized microbial cells of *Bacillus cereus*, *Bifidobacterium adolescentis*, *Clostridium beijerinckii*, *Deinococcus rediodurans*, *Enterococcus faecalis*, *Escherichia coli*, *Lactobacillus gasseri*, *Rhodobacter sphaeroides*, *Staphylococcus epidermidis*, and *S. mutans*. Freeze-dried material was rehydrated with 1 ml of PBS (−) (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, and 1.76 mM KH_2PO_4) and stored at −80˚C in 100 μl aliquot.

RNA extraction used RNeasy PowerBiofilm kit (QIAGEN) following the manufacturer's instruction. Two 100 μl aliquots were used as starting material. Obtained RNA solutions were eluted with 50 μl of water and mixed into a single tube (100 μl of RNA solution). Obtained RNA concentration was measured with a Qubit RNA HS kit (ThermoFisher). DNA extraction was performed using a DNeasy PowerSoil kit (QIAGEN) by following the manufacturer's instruction. The DNA solutions obtained were eluted with 50 μl of water and mixed into a single tube (100 μl of DNA solution). The DNA concentration obtained was measured with a Qubit DNA HS kit (ThermoFisher).

Amplicon-seq analysis

DNA the mock microbial community was used for amplicon-seq analysis with 16S small subunit ribosomal RNA (ssrRNA) gene sequences. We selected two hypervariable target region V4 and V3–V4 for the analysis. Amplicon-seq libraries were constructed using the Illumina "16S Metagenomic Sequencing Library Preparation" protocol with some modifications. Briefly, PCR reaction used PCR enzyme "KOD plus" (TOYOBO) and recommended reaction conditions (1.5 mM $MgSO_4$, 0.2 mM dNTP, 1 unit/50 μl KOD plus, and 0.2 pmoles/μl primers). We used a single-step instead of the original two-step PCR procedure. Primers were designed for the V4 region [10] and V3–V4 region [21].

(Bac515F_D501: 5′- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ATA GCC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T GT GCC AGC MGC CGC GGT AA -3′, Bac806R_D701: 5′- CAA GCA GAA GAC GGC ATA CGA GAT CGA GTA ATG TGA CTG GAG TTC AGA CGT GTG CTC TTC CGA TCT GGA CTA CHV GGG TWT CTA AT -3′)

(BacV3_V4_F_D502: 5′- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TAG AGG CAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TCC TAC GGG NGG CWG CAG -3′, BacV3_V4_R_D702: 5′- CAA GCA GAA GAC GGC ATA CGA GAT TCT CCG GAG TGA CTG GAG TTC AGA CGT GTG CTC TTC CGA TCT GAC TAC HVG GGT ATC TAA TCC -3′) including TruSeqHT index and linker sequences. V4 target and V3–V4 target reactions were amplified as 98˚C for 2 min, 25 cycles of 98˚C for 15 s, 55˚C for 45 s, 68˚C for 1 min, and 68˚C for 6 min. Products were purified with AMpure magnetic beads following the manufacturer's instructions and then eluted with 50 μl of water.

Obtained PCR products were quantified by quantitative PCR (qPCR) by a KAPA Library Quantification Kit Illumina Platform (KAPA biosystems) following the manufacturer's instructions. A 2 nM pool was constructed based on quantification results. This pool was used for Illumina MiSeq sequencing with 5% PhiX spike-in and obtained 250 bp paired-end reads. Obtained reads were analyzed with the QIIME2 pipeline [22] with DADA2 [23] for quality control and taxonomic assignment with a naïve Bayes classifier for annotation [24]. Each target region was specified by primer sequences to train the naïve classifier with silva132_99.fna of the Silva database, release 132 [4]. Annotation was on taxonomy_7_levels.txt in the same database. Obtained sequences were deposited in DDBJ DRA, accession number DRA009985.

ARI-seq analysis

Obtained total RNA from the mock microbial community was used to construct a total RNA-seq sequencing library with a SMARTer stranded RNA-seq kit (Clonetech) following the manufacturer's instruction. We used 5.8 ng of RNA as starting material, PCR was repeated for 12 cycles, and final products were eluted by 10 μl of water. The obtained sequencing library was quantified with a KAPA Library Quantification Kit Illumina Platform following the manufacturer's instructions. Again, a 2 nM pool was constructed based on quantification results. The pool was used for Illumina MiSeq sequencing with 5% PhiX spike-in and obtained 250 bp paired-end reads. Obtained sequences were deposited in DDBJ DRA, accession number DRA009985.

Obtained reads were trimmed by trimmomatic-0.39 [25] with option "ILLUMINACLIP: TruSeq_LT_HT.fa:5:30:7 MINLEN:100 HEADCROP:6 LEADING:20 TRAILING:20." PhiX sequences were removed by USEARCH 11.0.667 -filter_phix option [26]. Low complexity filtering was performed with USEARCH 11.0.667 -filter_lowc option [27]. Cleaned reads were used in the assembly process using Trinity v2.8.5 with a minimum_contig_length of 500 [28].

Cleaned reads were sorted into ssrRNA and non-ssrRNA reads using SortMeRNA with paired-in or paired-out options [27], respectively. Reference sequences for sorting with SortMeRNA were silva-arc-16s-id95.fasta, silva-bac-16s-id90.fasta, and silva-euk-18s-id95.fasta. Sorted ssrRNA reads were used for mapping against Trinity output (Trinity.fasta). Mapping was performed by bowtie2 v.2.3.5.1-linux-x86_64 with options -1 and -2 used to specify paired mapping mode, while option -U and forward and reverse reads were used to specify non-paired mapping mode. Finally, we used the bowtie2 process in "local mode." Resulting SAM files were transformed with samtools into BAM files and sorted. Sorted BAM files were used to obtain counting information by eXpress v.1.5.1-linux_x86_64 [29]. Count data truncated with a custom script to remove reads with fewer than 10 counts.

Annotation for ssrRNA data–query for extracted sequences from "Trinity.fasta" mapped with ssrRNA reads by SortMeRNA–used the QIIME2 feature classifier command [30] in three modes: (1) classify-sklearn (the same method used for amplicon-seq analysis with naïve Bayes classifier that trained by silva132_99.fna of the Silva release 132 database without region specification) [24], (2) classify-consensus-BLAST [consensus taxonomic assignment by BLAST+ (Bl),

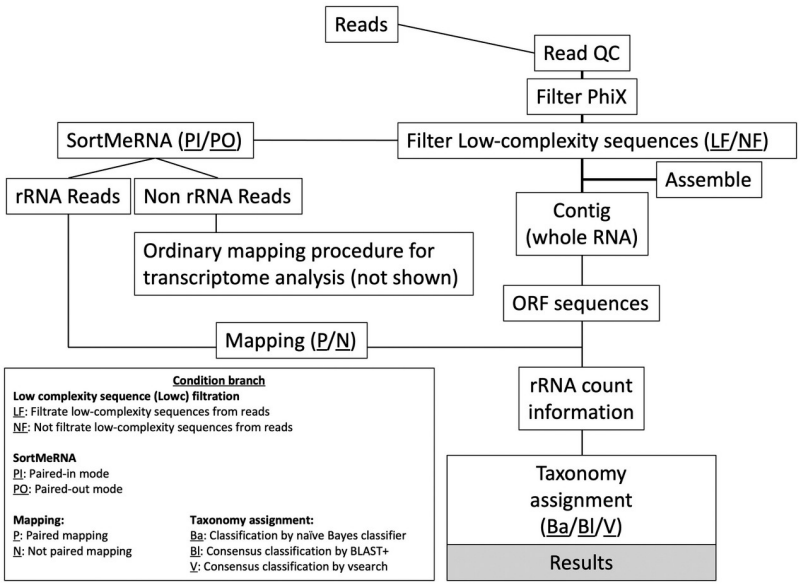


Fig 1. Analysis scheme. Flow chart of analysis process of mapping-based RNA-seq analysis to determine microbial community structure. Box indicates branching points in analysis conditions.

https://doi.org/10.1371/journal.pone.0254556.g001

first 10 hits] [31], and (3) classify-consensus-vsearch [consensus taxonomic assignment by vsearch (V), top 10 hits] [32].

Count data and annotation information were combined using an in-house script in R statistics software. Finally, we calculated reads per kilobase fragment (rpk) on the basis count data and query sequence length and then calculated relative abundance manually. The analysis scheme is illustrated in Fig 1, and analysis conditions are provided in S1 Table. Log of all scripts and commands will be provided upon request.

Search conditions for the ARI-seq approach and visualization of results

We used several different conditions for the four steps in the analysis pipeline (S1 Table, explanation of condition branch). First, we used two conditions in the reads qualification step. After trimming and artificial sequence removal, we added a low complexity sequence filtering step. A low complexity sequence is defined as a single nucleotide or short motif repeat in a read that can add noise to the assembly process. However, removal of low complexity sequences mainly affects short reads and can disrupt the assembly of reads to contigs. Therefore, we included the options to perform low complexity sequence filtering (LF) or not (NF).

Second, qualified reads were used for sorting to ssrRNA or non-ssrRNA sequences by SortMeRNA, which looks at forward and reverse reads individually; however, result output was paired reads (a set of forward read and reverse read). Hence, two strategies, "paired-in" and "paired-out," in the SortMeRNA program, can be used in the analysis. While a part of paired read was assigned as ssrRNA, the other part was assigned as non-ssrRNA. Both reads (= paired read) were assigned as ssrRNA in "paired-in (PI)" mode and both reads (= paired read) were assigned as non-ssrRNA in "paired-out (PO)" mode. These conditions affect numbers of reads in ssrRNA or non-ssrRNA categories and alter mapping results. Sorted ssrRNA reads were mapped into contigs by bowtie2.

Third, we examined "paired mapping (P)" and "non-paired mapping (N)." Normally, paired reads are used as mate pairs for mapping onto reference sequences (paired mapping).

However, bacterial RNA contains several different gene units as operons. We can expect that half of paired reads can be assigned on reference sequences. Paired reads should be separated into single reads and mapped separately (non-paired mapping). We provided options for these two strategies since they greatly affect mapping.

Reads assembled into contigs are used for taxonomic assignment by a part of the QIIME2 pipeline. Normally, a trained Bayes classifier is used for taxonomic assignment. However, we observed that results are not reliable using this approach. Thus, we included search options with a naïve Bayes classifier using consensus taxonomy classification (Bayes, Ba) with Bl, first 10 hits or consensus classification with V, top 10 hits. "Bl" and "V" are homology search base methods and show better performance than "Bayes (Ba)" conditions, as further discussed in the following section.

Words in parentheses indicate conditions options. Single and combinations of these options represent analysis conditions in the following sections. All possible analysis modes and its abbreviations are shown in S1 Table.

Obtained data were transformed to relative abundance and basic statistical values, such as total relative abundance value of all mock community member and average relative abundance of each mock member. These results are the basis for a cumulative bar plot. Distributions of relative abundance values were visualized with beeswarm plots and box plots. Statistics were calculated with R software. Boxplots show a whisker range of 1.5 × interquartile range and boxes that include first to third quartiles.

Comparison with other mapping-based RNA-seq analysis

Comparison between our method and the already reported mapping-based RNA-seq analysis was performed with meta-total RNA sequencing (MeTRS) technology [17]. First, we used MeTRS with our mock sequencing data to compare with our method results. Second, we obtained microbiome sequencing data to test MeTRS (SRR5439729 from the SRA database in GenBank) and analyzed it with both our method and MeTRS. MeTRS analysis was performed according to a study [17] with their scripts (https://github.com/normanpavelka/MeTRS) with Silva release 132 ssrRNA database. Some pipeline steps were slightly modified according to issue comments on the GitHub website (https://github.com/normanpavelka/MeTRS/issues/1). Revised codes and the resulting raw data will be provided upon request.

Results

Accuracy of taxonomic annotation

Amplicon-seq identified mock community members with high accuracy. Relative abundance (Fig 2) of 99.85% (V3–V4) to 99.90% (V4) for clustered fragments using QIIME2 are assigned correctly to genus. ARI-seq results showed contrasting results among three taxonomic assignment methods. Taxonomic assignment using a naïve Bayes classifier showed low accuracy. Depending on analysis conditions, the relative abundance of mock member genus assignments was only 22.17 ± 16.57%. Especially, "LF–PO mode ssrRNA sequence sorting" and "NF–PI mode ssrRNA sequence sorting" conditions showed very low accuracy (6.96 ± 3.25%, relative abundance of mock member genus). Other approaches correctly showed relative abundance to genus for 37.52% ± 2.70% of community members. Homology search methods (Bl and V) showed relatively high accuracy (94.31% ± 3.49%, relative abundance of mock member genera). MeTRS with our mock community data showed similar accuracy against homology search methods (98.20%, relative abundance of mock member genera).

Taxonomic assignment to "non-mock member" among analysis conditions (S2 Table) indicated that the ARI-seq approach with homology-based taxonomic assignment gave reasonable

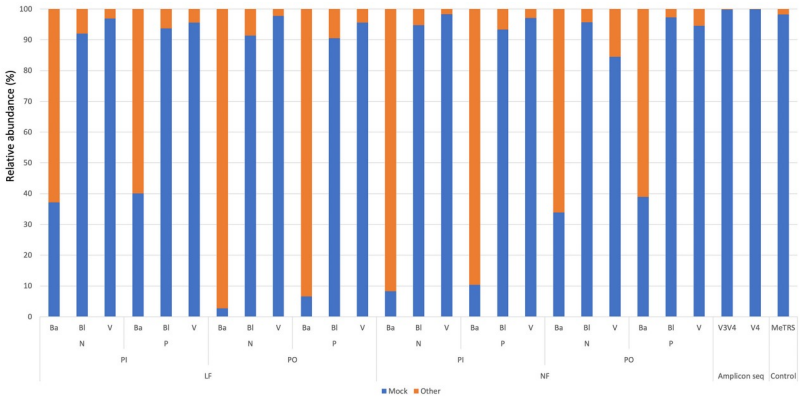


Fig 2. Accuracy of mock detection among tested methods. Blue bar indicates a detection rate of mock and orange bar indicates misdetection. Abbreviations of analysis condition in sample names are defined in S1 Table and in the main text.

https://doi.org/10.1371/journal.pone.0254556.g002

results, even considering relative abundance chart indications of non-mock member ssrRNA detection. Amplicon-seq detected small amounts of "non-mock members" and detected microbes with no taxonomical relationship with mock members. Homology-based taxonomic assignment of ARI-seq detected few such taxonomically independent sequences, possibly as contaminants (small amounts of *Homo sapience* 18S ssrRNA homolog, and human epidermal bacterium, *Enhydrobacter*, 16S ssrRNA homolog). Furthermore, most detected sequences by homology-based taxonomic assignments for ARI-seq are consensus sequences among kingdom, phylum, order, and family and include mock community members. This finding may reflect conserved regions of ssrRNA sequences that are shared broadly across taxonomically related genera of mock members. For example, genus, *Salmonella*, was detected. This species is closely related to genera, *Escherichia* and *Shigella*. In this context, such results do not indicate miss-assignment. The only exception is detection of plant chloroplast 16S in a few cases; however, detected amounts were low.

Taxonomic assignments by the naïve Bayes classifier for ARI-seq showed many false alignments. The conserved region of ssrRNA may be a problematic identifier. ARI-seq with homology-based taxonomy produced appropriate results compared with amplicon-seq findings.

Mapping-based total RNA-seq analysis for ssrRNA shows better mock community reconstruction

Relative abundance charts across analysis condition are presented in Fig 3. Except for ARI-seq taxonomy by a naïve Bayes classifier, all analysis conditions accurately detected all ten mock members (also see S3 Table). Amplicon-seq patterns can be uneven, and our results also showed such a pattern. Amplicon-seq with V3–V4 regions showed a significant abundance of *Escherichia–Shigella* and lower abundance of *Bifidobacterium*, *Enterococcus*, *Lactobacillus*, and *Staphylococcus*. Amplicon-seq with V4 region was somewhat less uneven than the V3–V4 amplicon. However, the abundance of *Bifidobacterium*, *Lactobacillus*, and *Staphylococcus* was quite small. In addition, MeTRS showed an uneven pattern, i.e., the pattern was different with amplicon-seq, and it showed a significant abundance of *Bacillus* and a lower abundance of *Bifidobacterium*, *Clostridium*, and *Lactobacillus*. Interestingly, some ARI-seq with homology-based taxonomic assignment showed more likely community structures than amplicon-seq. For example, for "NF–PI mode ssrRNA sequence sorting" with homology-based taxonomic assignment (Bl and V) and for both paired (P) and non-paired (N) mapping modes,

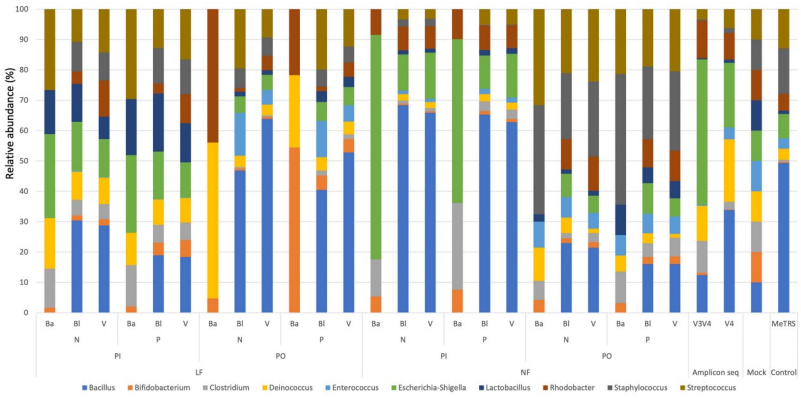


Fig 3. Accumulative bar chart of relative abundance among detected mock community members. Color chart is provided in the figure, and abbreviations for analysis methods are defined in S1 Table and in the main text.

https://doi.org/10.1371/journal.pone.0254556.g003

abundance pattern is quite even except for a very small abundance of *Enterococcus*. Furthermore, in "NF–PO mode ssrRNA sequence sorting" with homology-based taxonomic assignment (Bl and V) and both paired (P) and non-paired (N) mapping modes, abundance of *Bacillus*, *Staphylococcus*, and *Streptococcus* are relatively large, but a more even pattern is observed than for amplicon-seq and MeTRS results.

Distribution of abundance estimates is provided, as plotted in Fig 4. In "LF–PI mode ssrRNA sequence sorting" mode with homology-based taxonomic assignment (Bl ad V) and both paired (P) and non-paired (N) mapping modes, median abundance of mock members is almost 10%. Amplicon-sequence results showed median abundance of less than 10%, and the distribution of abundance estimates was broader than for ARI-seq. MeTRS showed a similar result with amplicon-seq. The median abundance of MeTRS was similar to that of V4 primer set, and the distribution of abundance was similar to that of V3V4 primer set. Thus, ARI-seq with "LF–PI mode ssrRNA sequence sorting" with homology-based taxonomic assignment (Bl and V) show better reconstruction performance for mock community structure than the amplicon-seq analysis pipeline.

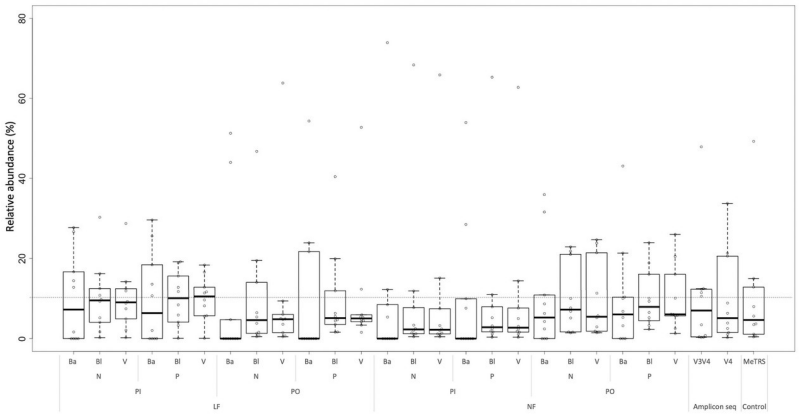
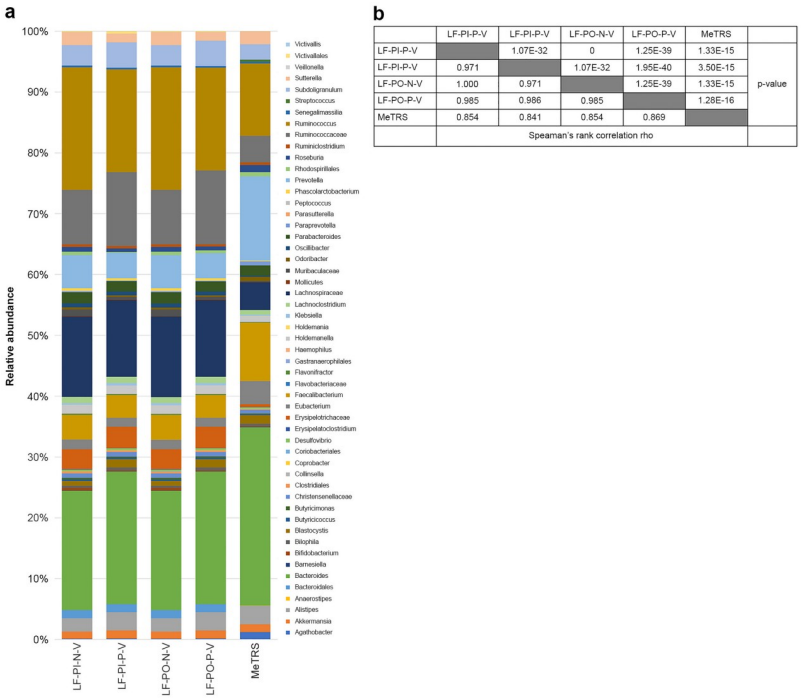


Fig 4. Scatter and box plots of distribution for relative abundance among mock community members. The figure shows scatter and box plots. Broken line indicates the 10% line of relative abundance expected from the fraction of each member in the original mock community. Abbreviations of analysis methods are provided in S1 Table and in the main text.

https://doi.org/10.1371/journal.pone.0254556.g004



	LF-PI-P-V	LF-PI-P-V	LF-PO-N-V	LF-PO-P-V	MeTRS	
LF-PI-P-V		1.07E-32	0	1.25E-39	1.33E-15	p-value
LF-PI-P-V	0.971		1.07E-32	1.95E-40	3.50E-15	
LF-PO-N-V	1.000	0.971		1.25E-39	1.33E-15	
LF-PO-P-V	0.985	0.986	0.985		1.28E-16	
MeTRS	0.854	0.841	0.854	0.869		
	Speaman's rank correlation rho					

Fig 5. Genera distribution in our method and MeTRS. (a) Relative abundance of the 53 genera commonly detected by our method (LF-PI-N-V, LF-PI-P-V, LF-PO-N-V and LF-PO-P-V) and MeTRS in the SRR5439729 data originated from a stool sample. (b) Spearman's rank correlation and *P*-value among the tested methods. Abbreviations of analysis condition in sample names are defined in S1 Table and in the main text.

https://doi.org/10.1371/journal.pone.0254556.g005

To test our method with "real-world" data, comparable analysis between our method and MeTRS was performed using published data from a human stool sample. We used SRA data published with MeTRS (SRR5439729) as basal stool microbiome data for this purpose. As shown in Fig 5, the composition of 53 genera commonly detected in this data by our method (LF-PI-N-V, LF-PI-P-V, LF-PO-N-V, and LF-PO-P-V modes) and MeTRS was similar to each other. Spearman's rank correlation and *P*-value indicated that those patterns are significantly similar.

Discussion

Our ARI-seq approach analysis of microbial populations shows genus-level annotation accuracy and reasonable quantitation among a mix of ten species in a mock community. The traditional total RNA-seq analysis pipeline using the "ribo-tag" concept displays limited taxonomic annotation (class level) [8], and recent work improves annotation only to order or family levels [15, 20]. Our mapping-based method with homology-based annotation showed genus-level accuracy with minor miss-mapping possible in conserved regions (S1 Table).

Results show that our method produces more precise quantitative data than amplicon-seq. Reconstruction of a mock community with ten bacterial species was optimal using (a) "LF–PI mode ssrRNA sequence sorting" with (b) homology-based taxonomic assignment (Bl and V) and (c) both paired (P) and non-paired (N) mapping modes. These features are commonly observed with total RNA-seq methods, and mock analyses using total RNA sequences showed similar results [16, 17, 20]. Indeed, the comparison between our method and MeTRS indicated that some of our analysis conditions showed better results than MeTRS as mock community reconstruction. Furthermore, "real-world" data trial showed that significant similar

community composition was reconstructed from stool RNA-seq data with both our method and MeTRS.

In conclusion, simple mapping-based quantification using ARI-seq displayed better performance for microbiome community reconstruction than amplicon-seq using specific analysis conditions. We optimized our ARI-seq approach by examining four factors in the analysis pipeline–LF, ssrRNA sequence sorting strategy, mapping strategy, and taxonomic assignment methods. Results indicate that removal of low complexity sequences (LF mode), sorting ssrRNA using paired-in mode (PI mode), and using homology-based taxonomic assignment (Bl and V mode) provide optimal reconstruction of a mock community. Total RNA-seq is widely used for meta-transcriptome analysis. The present study indicates that almost the same process can be used for microbiome analysis. Our process should open new opportunities for understanding functional microbiomes with a simple mapping-base analysis pipeline.

Supporting information

S1 Table. Branching points for analysis conditions. Four branching points in the analysis process, with abbreviations of analysis conditions in sample names.
(XLSX)

S2 Table. False positive detection. A list of false positive signals and abundance.
(XLSX)

S3 Table. Quantitative data for mock members. Reads per kilobase fragment is indicated in RNA-seq data (bundle column of non-paired mapping and paired mapping) and read counts in amplicon-seq data. These data are original data used to calculate relative abundance.
(XLSX)

Acknowledgments

A Linux-based computational platform used to analyze all data in this work was kindly provided by Yuichi Hongoh (Tokyo Institute of Technology). The author would like to thank Enago (www.enago.jp) for the English language review.

Author Contributions

Conceptualization: Shigeharu Moriya.

Data curation: Shigeharu Moriya.

Formal analysis: Shigeharu Moriya.

Investigation: Shigeharu Moriya.

Methodology: Shigeharu Moriya.

Project administration: Shigeharu Moriya.

Software: Shigeharu Moriya.

Supervision: Shigeharu Moriya.

Validation: Shigeharu Moriya.

Visualization: Shigeharu Moriya.

Writing – original draft: Shigeharu Moriya.

Writing – review & editing: Shigeharu Moriya.

References

1. Woese CR. Bacterial evolution. Microbiol Rev. 1987; 51: 221–271. https://doi.org/10.1128/mr.51.2.221-271.1987 PMID: 2439888
2. Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011; 166: 99–110. https://doi.org/10.1016/j.micres.2010.02.003 PMID: 20223646
3. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014; 42: D633–D642. https://doi.org/10.1093/nar/gkt1244 PMID: 24288368
4. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013; 41: D590–D596. https://doi.org/10.1093/nar/gks1219 PMID: 23193283
5. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006; 72: 5069–5072. https://doi.org/10.1128/AEM.03006-05 PMID: 16820507
6. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol. 1998; 180: 4765–4774. https://doi.org/10.1128/JB.180.18.4765-4774.1998 PMID: 9733676
7. Marchesi JR, Sato T, Weightman AJ, Martin TA, Fry JC, Hiom SJ, et al. Design and evaluation of useful Bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. Appl Environ Microbiol. 1998; 64: 795–799. https://doi.org/10.1128/AEM.64.2.795-799.1998 PMID: 9464425
8. Sogin ML, Morrison HG, Huber JA, Mark Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA. 2006; 103: 12115–12120. https://doi.org/10.1073/pnas.0605127103 PMID: 16880384
9. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011; 108;Suppl 1: 4516–4522. https://doi.org/10.1073/pnas.1000080107 PMID: 20534432
10. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012; 6: 1621–1624. https://doi.org/10.1038/ismej.2012.8 PMID: 22402401
11. Tveit AT, Urich T, Svenning MM. Metatranscriptomic analysis of arctic peat soil microbiota. Appl Environ Microbiol. 2014; 80: 5761–5772. https://doi.org/10.1128/AEM.01030-14 PMID: 25015892
12. Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLOS One. 2008; 3: e2527. https://doi.org/10.1371/journal.pone.0002527 PMID: 18575584
13. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol. 2014; 16: 2659–2671. https://doi.org/10.1111/1462-2920.12250 PMID: 24102695
14. Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. 2012; 13: R23. https://doi.org/10.1186/gb-2012-13-3-r23 PMID: 22455878
15. Söllinger A, Tveit AT, Poulsen M, Noel SJ, Bengtsson M, Bernhardt J, et al. Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. mSystems. 2018; 3: 39–19. https://doi.org/10.1128/mSystems.00038-18 PMID: 30116788
16. Li F, Henderson G, Sun X, Cox F, Janssen PH, Guan LL. Taxonomic assessment of rumen microbiota using total RNA and targeted amplicon sequencing approaches. Front Microbiol. 2016; 7: 987. https://doi.org/10.3389/fmicb.2016.00987 PMID: 27446027
17. Cottier F, Srinivasan KG, Yurieva M, Liao W, Poidinger M, Zolezzi F, et al. Advantages of meta-total RNA sequencing (MeTRS) over shotgun metagenomics and amplicon-based sequencing in the profiling of complex microbial communities. NPJ Biofilms Microbiomes. 2018; 4: 1–7.
18. Bang-Andreasen T, Anwar MZ, Lanzén A, Kjøller R, Rønn R, Ekelund F, et al. Total RNA sequencing reveals multilevel microbial community changes and functional responses to wood ash application in agricultural and forest soil. FEMS Microbiol Ecol. 2020: 96: fiaa016. https://doi.org/10.1093/femsec/fiaa016 PMID: 32009159
19. Tsuboi A, Itoga M, Hongoh Y, Moriya S. Mapping-based all-RNA-information sequencing analysis (ARIseq) pipeline simultaneously revealed taxonomic composition, gene expression, and their correlation in an acidic stream ecosystem. BioRXiv. 2017; 4: 1–29. https://doi.org/10.1101/159293

20. Yan YW, Zou B, Zhu T, Hozzein WN, Quan ZX. Modified RNA-seq method for microbial community and diversity analysis using rRNA in different types of environmental samples. PLOS One. 2017; 12: e0186161. https://doi.org/10.1371/journal.pone.0186161 PMID: 29016661

21. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013; 41: e1–e1. https://doi.org/10.1093/nar/gks808 PMID: 22933715

22. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. Nat Biotechnol. 2019; 37: 852–857.

23. Callahan BJ, Mcmurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016; 13: 581–583. https://doi.org/10.1038/nmeth.3869 PMID: 27214047

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12: 2825–2830.

25. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

26. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26: 2460–2461. https://doi.org/10.1093/bioinformatics/btq461 PMID: 20709691

27. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012; 28: 3211–3217. https://doi.org/10.1093/bioinformatics/bts611 PMID: 23071270

28. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29: 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440

29. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011; 12: R22. https://doi.org/10.1186/gb-2011-12-3-r22 PMID: 21410973

30. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018; 6: 90. https://doi.org/10.1186/s40168-018-0470-z PMID: 29773078

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009; 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

32. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016; 4: e2584. https://doi.org/10.7717/peerj.2584 PMID: 27781170