microorganisms

Communication

Optimization of Mapping Tools and Investigation of Ribosomal RNA Influence for Data-Driven Gene Expression Analysis in Complex Microbiomes

Ryo Mameda [1] and Hidemasa Bono [1,2,*]

[1] Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima 739-0046, Japan
[2] Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima 739-0046, Japan
* Correspondence: bonohu@hiroshima-u.ac.jp

Abstract: For gene expression analysis in complex microbiomes, utilizing both metagenomic and metatranscriptomic reads from the same sample enables advanced functional analysis. Due to their diversity, metagenomic contigs are often used as reference sequences instead of complete genomes. However, studies optimizing mapping strategies for both read types remain limited. In addition, although transcripts per million (TPM) is commonly used for normalization, few studies have evaluated the influence of ribosomal RNA (rRNA) in metatranscriptomic reads. This study compared Burrows–Wheeler Aligner–Maximal Exact Match (BWA-MEM) and Bowtie2 as mapping tools for metagenomic contigs. Even after optimizing Bowtie2 parameters, BWA-MEM showed higher efficiency in mapping both metagenomic and metatranscriptomic reads. Further analysis revealed that rRNA sequences contaminate predicted protein-coding regions in metagenomic contigs. When comparing TPM values across samples, contamination by rRNA led to an overestimation of TPM changes. This effect was more pronounced when the difference in rRNA content between samples was larger. These findings suggest that metatranscriptomic reads mapped to rRNA should be excluded before TPM calculations. This study highlights key factors influencing read mapping and quantification in gene expression analysis of complex microbiomes. The findings provide insights for improving analytical accuracy and advancing functional studies using both metagenomic and metatranscriptomic data.

Keywords: metagenomics; metatranscriptomics; NGS; gene expression; read mapping; ribosomal RNA

Academic Editor: Rajeev K. Azad

Received: 27 March 2025
Revised: 22 April 2025
Accepted: 25 April 2025
Published: 26 April 2025

Citation: Mameda, R.; Bono, H. Optimization of Mapping Tools and Investigation of Ribosomal RNA Influence for Data-Driven Gene Expression Analysis in Complex Microbiomes. *Microorganisms* **2025**, *13*, 995. https://doi.org/10.3390/microorganisms13050995

1. Introduction

Complex microbiomes, such as soil microbes, aquatic microbes, and gut bacteria, are ubiquitous in the environment and play critical roles in influencing the higher organisms with which they coexist. Additionally, certain complex microbiomes, such as those in activated sludge, are utilized in industrial applications, making the elucidation of their functions highly valuable.

Metagenomic and metatranscriptomic analyses using next-generation sequencing (NGS) have been widely employed to investigate the functions of complex microbiomes. Studies targeting individual microbial species include investigations of nitrification genes in *Nitrosomonas* sp. within activated sludge [1] and the discovery of antibiotic biosynthesis genes in Actinomycetes from soil microbiomes [2]. Broader processes, such as the nitrogen cycle [3] and carbon cycle [4], have also been analyzed in soil microbiomes. These examples underscore the increasing application of metagenomic and metatranscriptomic approaches in functional microbiome research.

Reference sequences is a part of critical factor in the functional analysis of complex microbiomes. In NGS-based analyses, NGS reads must be compared against reference sequences with functional annotations. For single organisms, complete genome sequences can serve as references; however, this approach is challenging for complex environmental microbiomes due to their high diversity. Recent advancements in microbial genome sequencing have introduced methods such as metagenome-assembled genomes (MAGs) for bacteria with relative abundances of less than 1% in complex microbiomes [5] and single-amplified genomes (SAGs) obtained through single-cell analysis [6]. Notably, approximately half of the genomes in Release 220 of the Genome Taxonomy Database consist of MAGs or SAGs [7]. Despite these advances, only an estimated 2.1% of environmental bacteria have been genome-sequenced [8], highlighting the limitations of using database-registered genome data as references for species and gene function identification in complex microbiomes. In previous studies, metagenomic contigs have been used as reference sequences and allow for more comprehensive coverage of reads obtained from samples containing a large proportion of uncultured microorganisms, such as soil microbiomes [9,10]. In addition, covering both metagenomic and metatranscriptomic reads enables more advanced differential gene expression analysis [11,12]. In such analyses, mapping reads to reference sequences and quantifying gene expression using read counts are common [9,10,13]. Some studies have focused on optimizing mapping tools specifically for metagenomic analysis, and Burrows–Wheeler Aligner (BWA) and Bowtie2 have demonstrated great benchmarking performance [14–16]. These mapping tools have consistently demonstrated efficient performance for metagenomic reads in previous studies. Although efficient mapping tools that can accurately align both types of reads are essential, there has been a lack of comparative analyses evaluating their effectiveness for metatranscriptomic reads.

In addition to read mapping, methods for evaluating gene expression levels must also be considered. Normalization is considered essential for comparing samples, and transcripts per million (TPM) is widely used [17,18]. TPM represents the proportion of read counts for a given gene relative to the total expressed genes in a sample, normalized by gene length. As a result, TPM values can be influenced by the expression levels of other genes present in the transcriptome. However, its specific effects on gene expression analysis in complex microbiomes have not been thoroughly investigated. This study focused on two key aspects of the gene expression analysis of complex microbiomes: read mapping and the influence of ribosomal RNA (rRNA) on TPM calculation. For read mapping, the widely used tools Burrows–Wheeler Aligner–Maximal Exact Matches (BWA-MEM) and Bowtie2 were compared. Additionally, contamination of rRNA sequences within the protein-coding sequences of metagenome contigs was detected, demonstrating its potential impact on TPM calculation. The analysis was conducted using metagenomic and metatranscriptomic reads obtained from the same samples available in the National Center for Biotechnology Information (NCBI) database.

2. Materials and Methods

The several shell scripts used in this research are available in the GitHub repository at https://github.com/RyoMameda/workflow (accessed on 22 April 2025). The development into a workflow language will be released.

2.1. Computational Resources

This analysis was conducted using a system equipped with 64 GB RAM and an Apple M1 Max chip, operating on macOS Sequoia. The binning step was performed on a Linux environment using the NIG supercomputer at the ROIS National Institute of Genetics.

2.2. Optimization of Mapping Tools

The analysis utilized 56 short-read datasets of soil microbiomes. Metagenomic and metatranscriptomic reads were obtained from 24 samples each, as described in reference studies [9,10,13,19] (Table S1). Two reference studies mentioned rRNA depletion before sequencing [9,13]. Over 95% of bases in the raw reads had a quality score of Q20 or higher, as confirmed by quality control described below. FASTQ files for these reads were retrieved from the NCBI Sequence Read Archive (SRA) database using the SRA Toolkit [20] (v3.0.10) with the prefetch and fasterq-dump commands. Quality control and trimming were performed using fastp [21] (v0.23.4) with the parameters -q 20 -t 1 -T 1. Trimmed metagenomic reads were assembled into contigs using MEGAHIT [22,23] (v1.2.6) with default parameters. Protein-coding sequences were predicted from metagenomic contigs using Prodigal [24] (v2.6.3) with the -p meta parameter. Trimmed metagenomic and metatranscriptomic reads were mapped to the predicted protein-coding sequences using BWA MEM [25] (v0.7.17) or Bowtie2 [26–28] (v2.5.1). Sequence Alignment/Map (SAM) files generated during mapping were converted to Binary Alignment/Map (BAM) format using SAMtools [29] (v1.17) with the sort command, and mapping statistics were analyzed with the flagstat command. Mapping quality was analyzed using Qualimap2 [30] (v2.3).

2.3. Gene Expression Analysis

Gene annotation was assigned to the predicted protein-coding sequences as follows. Predicted nucleotide sequences of metagenomic contigs were screened against rRNA sequences from NCBI [31] using BLASTN [32] (v2.15.0) with an E-value threshold of 0.1. The remaining protein sequences were annotated by querying against the Swiss-Prot database of well-characterized proteins in UniProt [33] using DIAMOND [34] (v2.0.15) with BLASTP and an E-value threshold of 0.1. Sequences without hits in Swiss-Prot were further annotated using protein domain information from Pfam [35] via HMMER [36] (v3.3.2) with an E-value threshold of 0.1, and were parallelized using GNU Parallel (v20230922) [37]. Predicted protein sequences were annotated based on the highest-ranking hits from the BLASTP and HMMER searches, with sequences lacking significant matches designated as hypothetical proteins. Using the annotations and BAM files from the previous steps, read counts for each sequence were quantified with the featureCounts command in Subread [38] (v2.0.6). Both metagenomic and metatranscriptomic read counts were normalized using the following equation. For a given contig t, let T_t represent the metatranscriptomic read count, and L_t the contig length. Transcripts per million (*TPM*) was calculated accordingly (1).

$$TPM = \frac{\left(\frac{T_\mathrm{t}}{L_\mathrm{t}}10^3\right)}{\sum_t \left(\frac{T_\mathrm{t}}{L_\mathrm{t}}10^3\right)}10^6 \tag{1}$$

3. Results and Discussion

3.1. Mapping Rates to Metagenomic Contigs

The analytical pipeline, encompassing processes from contig construction to gene expression quantification, was developed and integrated (Figure 1).

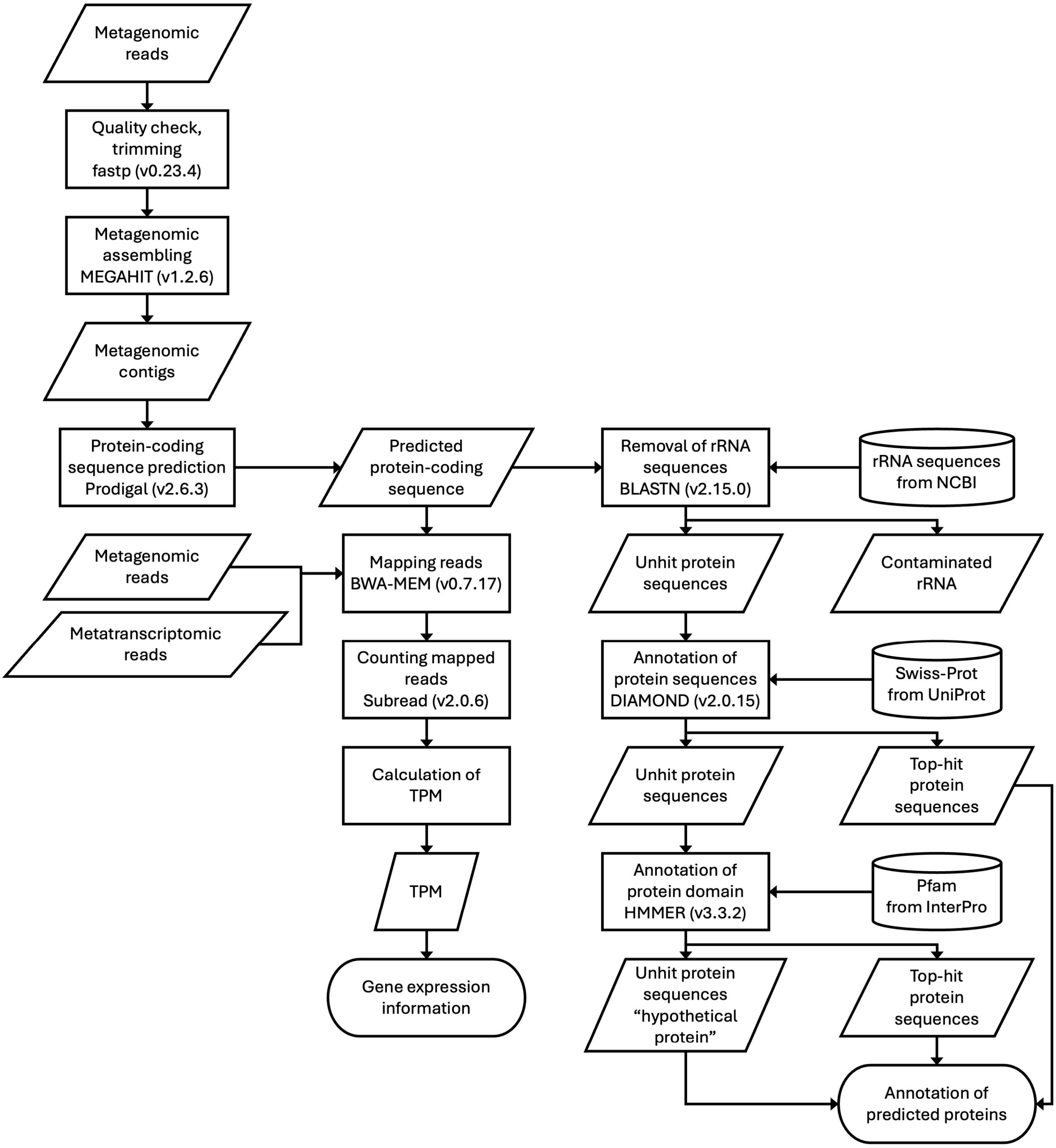
Metagenomic reads
Quality check, trimming fastp (v0.23.4)
Metagenomic assembling MEGAHIT (v1.2.6)
Metagenomic contigs
Protein-coding sequence prediction Prodigal (v2.6.3)
Predicted protein-coding sequence
Removal of rRNA sequences BLASTN (v2.15.0)
rRNA sequences from NCBI
Metagenomic reads
Metatranscriptomic reads
Mapping reads BWA-MEM (v0.7.17)
Unhit protein sequences
Contaminated rRNA
Counting mapped reads Subread (v2.0.6)
Annotation of protein sequences DIAMOND (v2.0.15)
Swiss-Prot from UniProt
Calculation of TPM
Unhit protein sequences
Top-hit protein sequences
TPM
Annotation of protein domain HMMER (v3.3.2)
Pfam from InterPro
Gene expression information
Unhit protein sequences “hypothetical protein”
Top-hit protein sequences
Annotation of predicted proteins

Figure 1. Data-driven analytical pipeline for gene expression analysis in complex microbiomes. Features of DNA sequences and read data are shown in rhombuses, processing methods are indicated in rectangles, datasets used as annotation references are placed in cylinders, and the final output data are presented in rounded rectangles.

First, both metagenomic and metatranscriptomic reads were mapped on metagenomic contigs constructed by MEGAHIT with BWA-MEM or Bowtie2 under default parameters. Overall, BWA-MEM achieved a higher mapping rate than Bowtie2 (sensitive preset) for both read types (Figure 2). Although previous studies reported that BWA-MEM achieves higher mapping rates for metagenomic contigs and those mapping rates were similar as the results in this study [14], it was hypothesized that Bowtie2 parameters might not have been fully optimized. While BWA-MEM allows local alignment and has a default seed length of 19, the Bowtie2-sensitive preset is end-to-end mode. To bring Bowtie2 settings closer to BWA-MEM, local or very-sensitive-local preset with a seed length of 19 (-L 19) was used. This local alignment setting significantly improved mapping rates (Figure 2). However, modifying Bowtie2's mismatch penalty to match BWA-MEM (setting it to 4) did not further improve mapping rates (Figure S1). Given these optimizations, the change in mapping rate was expected to affect mapping quality. Contrary to expectations, parameter changes did not result in noticeable differences in mapping quality (Figure S2). These results indicate that while optimizing Bowtie2 parameters can improve mapping rates, BWA-MEM remains a more efficient tool for mapping both metagenomic and metatranscriptomic reads. This finding represents an important insight for the gene expression analysis of complex microbiomes utilizing both types of reads. Both Bowtie2 and BWA-MEM utilize FM-index [39] and Burrows–Wheeler transform [40], but they differ in alignment features: BWA-MEM bases alignment on the number of mismatches, whereas Bowtie2 relies on an alignment score [41]. Whether this difference accounts for the observed variation in mapping rates remains unclear, requiring further investigation.

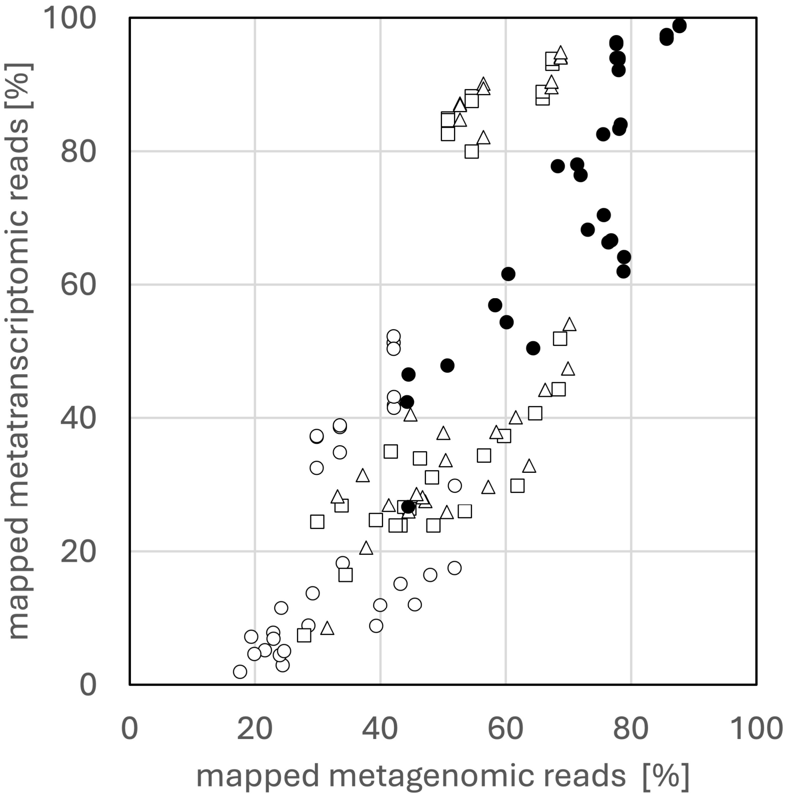

Figure 2. Mapping rates of metagenomic and metatranscriptomic reads. Reads were mapped using BWA-MEM (filled circle) or Bowtie2 (outlined shapes), whose setting was sensitive (circles), local -L 19 (squares), or very-sensitive-local -L 19 (triangles).

3.2. *Influence of rRNA on TPM Calculation*

Predicted protein sequences derived from the metagenomic contigs were found to include rRNA sequences due to chance translational frames. To address this, BLASTN searches were conducted against rRNA sequence databases from NCBI. For instance, in the metagenome contigs assembled from SRR24888648, only 0.09% of the predicted protein-coding sequences were identified as rRNA. However, an analysis of mapped read counts by BWA-MEM revealed that 0.16% (155,043 of 98,847,988) of metagenomic reads and 36.0% (23,079,523 of 64,026,082) of metatranscriptomic reads (SRR24887388) mapped to rRNA sequences. Although in vitro rRNA removal was performed in the referenced study [9], residual rRNA contamination persisted in metatranscriptomic reads, affecting predicted protein-coding sequences. In contrast, for the predicted protein-coding contigs of SRR22507541, in which rRNA was not removed in vitro [19], 95.1% (34,017,387 of 35,774,766) of the metatranscriptomic reads (SRR22506317) were mapped to rRNA. While rRNA removal kits are known to be effective, residual rRNA sequences are still commonly observed [42,43]. This study further revealed that such rRNA sequences can also contaminate predicted protein-coding sequences, highlighting an additional source of potential bias in downstream analyses.

The impact of residual rRNA on TPM calculation was investigated. In a reference dataset [9] where rRNA was removed in vitro, the total TPM value of rRNA-mapped reads was 305,514 and 536,616 in SRR24888495 and SRR24887388, respectively. For genes commonly expressed in both samples, TPM values were calculated with and without including rRNA, and changes were expressed as $\log_{10}$ fold differences. The results showed that TPM values calculated with rRNA included were slightly but consistently higher compared to those calculated with rRNA excluded (Figure 3A). In contrast, when comparing SRR24887388 with SRR24887267 (total rRNA TPM values are 536,616 and 513,379, respectively), the variability in gene expression between samples was minimal (Figure 3B). To further assess the impact, an extreme case was analyzed using data from another reference study [19] in which metatranscriptomic reads were obtained without in vitro rRNA removal. TPM values were calculated both with rRNA included and after the removal of rRNA. In SRR22506317 and SRR22506321, total rRNA TPM values were 970,933 and 980,113, respectively. When comparing gene expression ratios between the two samples, the difference became more noticeable when rRNA was excluded from only one of the samples during TPM calculation (Figure 3C). These findings indicate that larger differences in rRNA contamination between samples result in greater discrepancies in gene expression variability. Therefore, consistent rRNA removal across samples is essential during data processing to ensure accurate expression analysis.

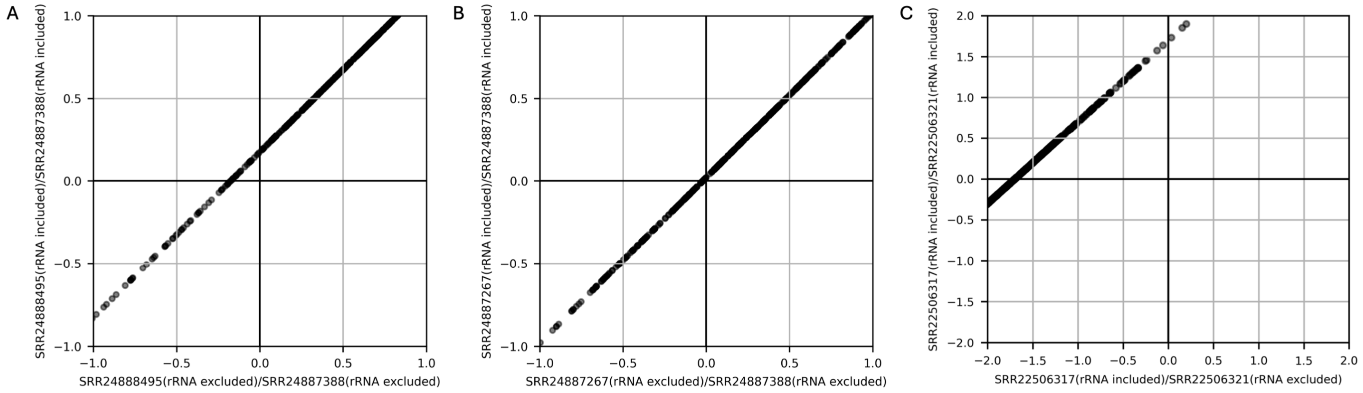

Figure 3. Log-fold change in TPM between samples. A random subset of 500 expressed genes were plotted. A comparison was conducted between (**A**) SRR24888495 and SRR24887388, (**B**) SRR24887267 and SRR24887388, (**C**) SRR22506317 and SRR22506321 to evaluate the impact of rRNA removal.

These results demonstrate that when the residual rRNA varies across samples, calculating TPM with rRNA included can lead to inconsistencies in the interpretation of differential gene expression analysis. By excluding rRNA read counts from TPM calculation, such inter-sample noise can be minimized, allowing for a more accurate assessment of gene expression variation. Although the potential impact of rRNA on TPM has been previously mentioned in single organisms [44], this study showed that differences in calculation can lead to measurable variations. Alternatively, tools such as SortMeRNA [45] can be employed to remove rRNA sequences in silico from metatranscriptomic reads prior to analysis. Further investigation is needed to evaluate the effectiveness of these approaches in detail. Additionally, TPM represents the relative expression level normalized across samples. In complex microbiomes, where genes originate from various microbial species, normalization methods using both read counts and gene length might improve the accuracy of gene-specific expression analysis [46]. Furthermore, incorporating the gene copy number from the metagenome into the analysis has been suggested to further refine the evaluation of gene expression dynamics [12]. Improving cross-sample quantification accuracy may require approaches such as incorporating internal standards during NGS read acquisition [47]. Furthermore, more efficient methods for extracting DNA and RNA from environmental microbiome samples need to be explored [48–50]. Combining such advanced sample preparation techniques with this study could further enhance the precision and scalability of quantitative gene expression analysis in complex microbiomes.

4. Conclusions

This study focused on the gene expression analysis of complex microbiomes using NGS. It is well established that mapping both metagenomic and metatranscriptomic reads obtained from the same sample to metagenomic contigs enables advanced expression analysis. Evaluation of mapping tools in this study revealed that BWA-MEM efficiently maps both types of reads. Additionally, TPM is commonly used for normalizing mapped read counts during gene expression analysis. However, rRNA contamination was found in predicted protein sequences derived from metagenomic contigs, and metatranscriptomic reads often contain residual rRNA that cannot be completely removed in vitro. These factors were shown to potentially influence TPM calculation. Applying these findings to the functional analysis of complex microbiomes can contribute to more accurate and advanced gene expression analysis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/microorganisms13050995/s1, Table S1: NGS reads of soil microbiome used in this study.; Figure S1: Mapping rates of metagenomic and metatranscriptomic reads using Bowtie2.; Figure S2: Mapping quality of metagenomic and metatranscriptomic reads using Bowtie2. Refs. [9,10,13,20] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, R.M.; methodology, R.M.; software, R.M.; validation, R.M.; formal analysis, R.M.; investigation, R.M.; resources, H.B.; data curation, R.M.; writing—original draft preparation, R.M.; writing—review and editing, H.B.; visualization, R.M.; supervision, H.B.; project administration, R.M. and H.B.; funding acquisition, H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Center of Innovation for Bio-Digital Transformation (BioDX), the open innovation platform for industry–academia co-creation (COI-NEXT), the Japan Science and Technology Agency (JST; Grant Number JPMJPF2010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The shell scripts used in this research are available in the GitHub repository at https://github.com/RyoMameda/workflow (accessed on 22 April 2025).

Acknowledgments: Computations were partially conducted on the NIG supercomputer at the ROIS National Institute of Genetics.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NGS	Next-Generation Sequencing
MAGs	Metagenome-Assembled Genomes
SAGs	Single-Amplified Genomes
TPM	Transcripts Per Million
rRNA	Ribosomal RNA
BWA-MEM	Burrows–Wheeler Aligner–Maximal Exact Matches
NCBI	National Center for Biotechnology Information
SRA	Sequence Read Archive
SAM	Sequence Alignment/Map
BAM	Binary Alignment/Map

References

1. Sato, Y.; Hori, T.; Koike, H.; Navarro, R.R.; Ogata, A.; Habe, H. Transcriptome Analysis of Activated Sludge Microbiomes Reveals an Unexpected Role of Minority Nitrifiers in Carbon Metabolism. *Commun. Biol.* **2019**, *2*, 179. [CrossRef] [PubMed]
2. Yang, S.; Zhang, W.; Yang, B.; Feng, X.; Li, Y.; Li, X.; Liu, Q. Metagenomic Evidence for Antibiotic-Associated Actinomycetes in the Karamay Gobi Region. *Front. Microbiol.* **2024**, *15*, 1330880. [CrossRef]
3. Nelson, M.B.; Martiny, A.C.; Martiny, J.B.H. Global Biogeography of Microbial Nitrogen-Cycling Traits in Soil. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 8033–8040. [CrossRef] [PubMed]
4. Hartman, W.H.; Ye, R.; Horwath, W.R.; Tringe, S.G. A Genomic Perspective on Stoichiometric Regulation of Soil Carbon Cycling. *ISME J.* **2017**, *11*, 2652–2665. [CrossRef] [PubMed]
5. Albertsen, M.; Hugenholtz, P.; Skarshewski, A.; Nielsen, K.L.; Tyson, G.W.; Nielsen, P.H. Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes. *Nat. Biotechnol.* **2013**, *31*, 533–538. [CrossRef] [PubMed]
6. Lasken, R.S.; McLean, J.S. Recent Advances in Genomic DNA Sequencing of Microbial Species from Single Cells. *Nat. Rev. Genet.* **2014**, *15*, 577–584. [CrossRef]
7. The University of Queenland. The Genome Taxonomy Database. Available online: https://gtdb.ecogenomic.org/ (accessed on 29 December 2024).
8. Zhang, Z.; Wang, J.; Wang, J.; Wang, J.; Li, Y. Estimate of the Sequenced Proportion of the Global Prokaryotic Genome. *Microbiome* **2020**, *8*, 134. [CrossRef]
9. Honeker, L.K.; Pugliese, G.; Ingrisch, J.; Fudyma, J.; Gil-Loaiza, J.; Carpenter, E.; Singer, E.; Hildebrand, G.; Shi, L.; Hoyt, D.W.; et al. Drought Re-Routes Soil Microbial Carbon Metabolism towards Emission of Volatile Metabolites in an Artificial Tropical Rainforest. *Nat. Microbiol.* **2023**, *8*, 1480–1494. [CrossRef]
10. Mendes, L.W.; Raaijmakers, J.M.; De Hollander, M.; Sepo, E.; Gómez Expósito, R.; Chiorato, A.F.; Mendes, R.; Tsai, S.M.; Carrión, V.J. Impact of the Fungal Pathogen Fusarium Oxysporum on the Taxonomic and Functional Diversity of the Common Bean Root Microbiome. *Environ. Microbiome* **2023**, *18*, 68. [CrossRef]
11. Yu, K.; Zhang, T. Metagenomic and Metatranscriptomic Analysis of Microbial Community Structure and Gene Expression of Activated Sludge. *PLoS ONE* **2012**, *7*, e38183. [CrossRef]
12. Zhang, Y.; Thompson, K.N.; Huttenhower, C.; Franzosa, E.A. Statistical Approaches for Differential Expression Analysis in Metatranscriptomics. *Bioinformatics* **2021**, *37*, i34–i41. [CrossRef]
13. Freeman, E.C.; Emilson, E.J.S.; Dittmar, T.; Braga, L.P.P.; Emilson, C.E.; Goldhammer, T.; Martineau, C.; Singer, G.; Tanentzap, A.J. Universal Microbial Reworking of Dissolved Organic Matter along Environmental Gradients. *Nat. Commun.* **2024**, *15*, 187. [CrossRef] [PubMed]
14. Jaillard, M.; Tournoud, M.; Meynier, F.; Veyrieras, J.-B. Optimization of Alignment-Based Methods for Taxonomic Binning of Metagenomics Reads. *Bioinformatics* **2016**, *32*, 1779–1787. [CrossRef] [PubMed]
15. Armstrong, G.; Martino, C.; Morris, J.; Khaleghi, B.; Kang, J.; DeReus, J.; Zhu, Q.; Roush, D.; McDonald, D.; Gonazlez, A.; et al. Swapping Metagenomics Preprocessing Pipeline Components Offers Speed and Sensitivity Increases. *mSystems* **2022**, *7*, e01378-21. [CrossRef]
16. Wang, L.; Ding, R.; He, S.; Wang, Q.; Zhou, Y. A Pipeline for Constructing Reference Genomes for Large Cohort-Specific Metagenome Compression. *Microorganisms* **2023**, *11*, 2560. [CrossRef] [PubMed]

17. Lopez-Fernandez, M.; Simone, D.; Wu, X.; Soler, L.; Nilsson, E.; Holmfeldt, K.; Lantz, H.; Bertilsson, S.; Dopson, M. Metatranscriptomes Reveal That All Three Domains of Life Are Active but Are Dominated by Bacteria in the Fennoscandian Crystalline Granitic Continental Deep Biosphere. *mBio* **2018**, *9*, e01792-18. [CrossRef]
18. Braga, L.P.P.; Pereira, R.V.; Martins, L.F.; Moura, L.M.S.; Sanchez, F.B.; Patané, J.S.L.; Da Silva, A.M.; Setubal, J.C. Genome-Resolved Metagenome and Metatranscriptome Analyses of Thermophilic Composting Reveal Key Bacterial Players and Their Metabolic Interactions. *BMC Genom.* **2021**, *22*, 652. [CrossRef]
19. Li, X.; Bei, Q.; Rabiei Nematabad, M.; Peng, J.; Liesack, W. Time-Shifted Expression of Acetoclastic and Methylotrophic Methanogenesis by a Single Methanosarcina Genomospecies Predominates the Methanogen Dynamics in Philippine Rice Field Soil. *Microbiome* **2024**, *12*, 39. [CrossRef]
20. SRA Toolkit Development Team. SRA Toolkit. Available online: https://github.com/ncbi/sra-tools (accessed on 29 December 2024).
21. Chen, S. Ultrafast One-pass FASTQ Data Preprocessing, Quality Control, and Deduplication Using Fastp. *iMeta* **2023**, *2*, e107. [CrossRef]
22. Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct *de. Bruijn* Graph. *Bioinformatics* **2015**, *31*, 1674–1676. [CrossRef]
23. Li, D.; Luo, R.; Liu, C.-M.; Leung, C.-M.; Ting, H.-F.; Sadakane, K.; Yamashita, H.; Lam, T.-W. MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven by Advanced Methodologies and Community Practices. *Methods* **2016**, *102*, 3–11. [CrossRef] [PubMed]
24. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
25. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
26. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef]
27. Langmead, B.; Salzberg, S.L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
28. Langmead, B.; Wilks, C.; Antonescu, V.; Charles, R. Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors. *Bioinformatics* **2019**, *35*, 421–432. [CrossRef] [PubMed]
29. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008. [CrossRef] [PubMed]
30. Okonechnikov, K.; Conesa, A.; García-Alcalde, F. Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data. *Bioinformatics* **2016**, *32*, 292–294. [CrossRef]
31. NCBI rRNA Sequences. Available online: https://ftp.ncbi.nlm.nih.gov/blast/db/ (accessed on 29 December 2024).
32. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and Applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
33. UniProt-EMBL-EBI Swiss-Prot Database. Available online: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz (accessed on 29 December 2024).
34. Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. *Nat. Methods* **2021**, *18*, 366–368. [CrossRef]
35. InterPro-EMBL-EBI Pfam Database. Available online: https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz (accessed on 29 December 2024).
36. HMMER. Available online: http://hmmer.org/ (accessed on 29 December 2024).
37. Tange, O. GNU Parallel 20230922 ('Derna'). Available online: https://zenodo.org/records/8374296 (accessed on 22 April 2025).
38. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. *Bioinformatics* **2014**, *30*, 923–930. [CrossRef]
39. Ferragina, P.; Manzini, G. Opportunistic Data Structures with Applications. In Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, USA, 12–14 November 2000; IEEE Computer Society: Washington, DC, USA, 2000; pp. 390–398.
40. Burrows, M.; Wheeler, D.J. A Block-Sorting Lossless Data Compression Algorithm 1994. Available online: https://www.cl.cam.ac.uk/teaching/2003/DSAlgs/SRC-124.pdf (accessed on 22 April 2025).
41. Hatem, A.; Bozdag, D.; Toland, A.E.; Çatalyürek, Ü.V. Benchmarking Short Sequence Mapping Tools. *BMC Bioinform.* **2013**, *14*, 184. [CrossRef] [PubMed]
42. Stewart, F.J.; Ottesen, E.A.; DeLong, E.F. Development and Quantitative Analyses of a Universal rRNA-Subtraction Protocol for Microbial Metatranscriptomics. *ISME J.* **2010**, *4*, 896–907. [CrossRef] [PubMed]
43. He, S.; Wurtzel, O.; Singh, K.; Froula, J.L.; Yilmaz, S.; Tringe, S.G.; Wang, Z.; Chen, F.; Lindquist, E.A.; Sorek, R.; et al. Validation of Two Ribosomal RNA Removal Methods for Microbial Metatranscriptomics. *Nat. Methods* **2010**, *7*, 807–812. [CrossRef]

44. Zhao, Y.; Li, M.-C.; Konaté, M.M.; Chen, L.; Das, B.; Karlovich, C.; Williams, P.M.; Evrard, Y.A.; Doroshow, J.H.; McShane, L.M.; et al. A Comparative Study of Quantification Measures for the Analysis of RNA-Seq Data from the NCI Patient-Derived Models Repository. *J. Transl. Med.* **2021**, *19*, 269. [CrossRef] [PubMed]
45. Kopylova, E.; Noé, L.; Touzet, H. SortMeRNA: Fast and Accurate Filtering of Ribosomal RNAs in Metatranscriptomic Data. *Bioinformatics* **2012**, *28*, 3211–3217. [CrossRef]
46. Klingenberg, H.; Meinicke, P. How to Normalize Metatranscriptomic Count Data for Differential Expression Analysis. *PeerJ* **2017**, *5*, e3859. [CrossRef]
47. Hardwick, S.A.; Chen, W.Y.; Wong, T.; Kanakamedala, B.S.; Deveson, I.W.; Ongley, S.E.; Santini, N.S.; Marcellin, E.; Smith, M.A.; Nielsen, L.K.; et al. Synthetic Microbe Communities Provide Internal Reference Standards for Metagenome Sequencing and Analysis. *Nat. Commun.* **2018**, *9*, 3096. [CrossRef]
48. Mori, H.; Kato, T.; Ozawa, H.; Sakamoto, M.; Murakami, T.; Taylor, T.D.; Toyoda, A.; Ohkuma, M.; Kurokawa, K.; Ohno, H. Assessment of Metagenomic Workflows Using a Newly Constructed Human Gut Microbiome Mock Community. *DNA Res.* **2023**, *30*, dsad010. [CrossRef]
49. Thorn, C.E.; Bergesch, C.; Joyce, A.; Sambrano, G.; McDonnell, K.; Brennan, F.; Heyer, R.; Benndorf, D.; Abram, F. A Robust, Cost-effective Method for DNA, RNA and Protein Co-extraction from Soil, Other Complex Microbiomes and Pure Cultures. *Mol. Ecol. Resour.* **2019**, *19*, 439–455. [CrossRef]
50. Shaffer, J.P.; Marotz, C.; Belda-Ferre, P.; Martino, C.; Wandro, S.; Estaki, M.; Salido, R.A.; Carpenter, C.S.; Zaramela, L.S.; Minich, J.J.; et al. A Comparison of DNA/RNA Extraction Protocols for High-Throughput Sequencing of Microbial Communities. *BioTechniques* **2021**, *70*, 149–159. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.