Received: 19 August 2022 | Accepted: 16 May 2023

DOI: 10.1111/2041-210X.14149

Methods in Ecology and Evolution

RESEARCH ARTICLE

Reconstruction of small subunit ribosomal RNA from high-throughput sequencing data: A comparative study of metagenomics and total RNA sequencing

Christopher A. Hempel[1,2] | Shea E. E. Carson[1] | Tyler A. Elliott[1] | Sarah J. Adamowicz[1] | Dirk Steinke[1,2]

[1]Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada

[2]Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada

Correspondence
Dirk Steinke
Email: dsteinke@uoguelph.ca

Funding information
Food from Thought: Agricultural Systems for a Healthy Planet Initiative; Canada First Research Excellence Fund, Grant/Award Number: 000054; Government of Canada, Grant/Award Number: 15401; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2022-04569

Handling Editor: Antonino Malacrinò

Abstract

1. The small subunit (SSU) ribosomal RNA (rRNA) is the most commonly used marker for the identification of microbial taxa, but its full-length reconstruction from high-throughput sequencing (HTS) data remains challenging. Metagenomics and total RNA sequencing (total RNA-Seq) are target-PCR-free HTS methods that are used to characterize microbial communities and simultaneously reconstruct SSU rRNA sequences. However, more testing is required to determine and improve their effectiveness.
2. We processed metagenomics and total RNA-Seq data retrieved from a commercially available mock microbial community and an aquarium sample using 112 combinations of data processing tools. We determined the SSU rRNA reconstruction completeness of both sequencing methods for both samples and analysed the impact of data processing tools on SSU rRNA completeness.
3. In contrast to metagenomics, total RNA-Seq allowed for the complete or near-complete reconstruction of all mock community SSU rRNA sequences and generated up to 438 SSU rRNA sequences with ≥80% completeness from the aquarium sample using only 1/5 of an Illumina MiSeq run. SSU rRNA completeness of metagenomics significantly correlated with the genome size of mock community species. Data processing tools impacted SSU rRNA completeness, in particular the utilized assemblers.
4. These results are promising for the high-throughput reconstruction of novel full-length SSU rRNA sequences and could advance the simultaneous application of multiple -omics approaches in routine environmental assessments to allow for more holistic assessments of ecosystems.

KEYWORDS
bioinformatics, data processing tool benchmarking, metagenomics, metatranscriptomics, microbial identification, mock community, small subunit ribosomal RNA, total RNA sequencing

1 | INTRODUCTION

Microbial organisms (prokaryotes and unicellular eukaryotes) make up most of the biodiversity on our planet, with the global number of microbial species estimated at up to one trillion (Locey & Lennon, 2016). These organisms play crucial roles within every biome on Earth, including the microbiomes within other organisms (e.g. humans). At the same time, they are very sensitive to environmental change. In fact, the microbial community composition of a particular environment can provide us with important information about its state and health (Cordier et al., 2019; Pawlowski et al., 2016; Proctor et al., 2019; Sagova-Mareckova et al., 2021; Smith et al., 2015).

Determining this community composition requires the identification of its member taxa. Since most microbes lack diagnostic traits and are too abundant and diverse to be identified morphologically, their identity is typically determined by using DNA-based methods (Pawlowski et al., 2012; Woese, 1987), which require reference databases. Ideally, these contain full-length genomes of as many as possible microbial taxa as possible to allow for taxonomic annotations at the highest possible resolution. However, most microbial taxa are unknown, and even for known taxa, the number of available full-length or even partial genomes is very low because the high-throughput reconstruction of full-length genomes has only recently become viable. As a consequence, the largest and most widely used curated database for genomes, NCBI RefSeq, only comprises genomes of 71,311 microbial organisms (release 213: https://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/archive/RefSeq-release213.catalog.gz). The largest project cataloguing all microbes on Earth (Earth Microbiome Project: http://www.earthmicrobiome.org) has led to the reconstruction of 52,515 microbial genomes to date (Nayfach et al., 2021). This clearly shows that the reconstruction of all microbial genomes or at least a larger segment of microbial diversity will require much more time.

For the time being, a more feasible and common approach to identifying microbial taxa is to focus on specific genes. The small subunit (SSU) ribosomal RNA (rRNA) gene (16S rRNA for prokaryotes and 18S rRNA for eukaryotes) is the primary marker gene for microbes. SSU rRNA reference databases are also far from complete, but the most commonly used database, SILVA, contains 510,508 full-length, non-redundant SSU rRNA gene sequences (release 138.1 SSU Ref NR 99: https://www.arb-silva.de/documentation/release-1381/), and this number is growing rapidly. Full-length SSU rRNA reference sequences are paramount because shorter sequences, which are generated by short-read amplicon sequencing, can lead to inaccurate taxonomic identifications (Johnson et al., 2019; Yarza et al., 2014). The traditional approach to reconstructing full-length SSU rRNA reference sequences involves cloning and Sanger sequencing, which is costly and comes with limited throughput. Advancements in sequencing technology gave rise to new approaches at lower costs and with higher throughput. Such high-throughput sequencing (HTS) approaches allow us to characterize the microbial community composition of many environmental samples simultaneously and are therefore extremely powerful to advance our understanding of microbial communities.

Multiple HTS approaches have been developed to date, and they can primarily be categorized into short-read sequencing (including synthetic long-read sequencing) and true long-read sequencing. True long-read sequencing reads have an average accuracy of >99.5% (Pacific Biosciences) or 92%–94% (Oxford Nanopore Technologies) at an average length of 10–25 kb (Delahaye & Nicolas, 2021; Hon et al., 2020), which is multiple times the length of the complete SSU rRNA (approximately 1.46 kb on average, based on the average length of all sequences in SILVA release 138.1 SSU Ref NR 99). However, although true long-read sequencing has been applied successfully for some microbial community analyses (Singer et al., 2016; Tedersoo et al., 2021), it is limited by the available read depth of long-read instruments. This makes it less scalable and currently not applicable for the study of complex microbial communities in comparison with short-read or synthetic long-read sequencing.

Metagenomics and metatranscriptomics are short-read sequencing methods without target PCR that are generally applied to analyse the presence and expression of functional genes within communities (Almeida & De Martinis, 2019; Bashiardes et al., 2016; Shakya et al., 2019; Wooley et al., 2010). Both involve random fragmentation of genomes and transcripts into shorter sequences and can thereby randomly cover smaller fragments of the SSU rRNA of organisms in a sample. These fragments can be utilized for SSU rRNA sequence reconstruction and the simultaneous taxonomic identification of communities. Metagenomics targets all DNA in a sample and results in broad genomic coverage, which is important e.g. for functional analyses of communities. However, this also means that the coverage of particular genes of interest, such as the SSU rRNA gene, is low. Consequently, SSU (and LSU) rRNA sequences can make up as little as 0.05%–1.4% of metagenomics sequences (Logares et al., 2014; Yilmaz et al., 2011). Targeted tools have been developed to extract and reconstruct SSU rRNA sequences from metagenomics datasets (Bengtsson-Palme et al., 2015; Fan et al., 2012; Miller et al., 2011; Pericard et al., 2018; Yuan et al., 2015; Zeng et al., 2017), but overall, the low proportion of SSU rRNA makes the approach inefficient for SSU rRNA reconstruction and taxonomic community analysis.

Metatranscriptomics typically uses only messenger RNA (mRNA), thereby providing a snapshot of the gene expression profile of a community. It also reveals which genes and metabolic pathways are active at a given time. As mRNA only makes up 1%–5% of cellular RNA (Milo & Phillips, 2015; Westermann et al., 2012), metatranscriptomics typically involves an mRNA enrichment step to increase the sequencing depth of mRNA. However, this means that rRNA, including SSU rRNA, is excluded from typical metatranscriptomics experiments. It is possible to skip the enrichment step and sequence all RNA instead, including rRNA. This approach has been referred to as double-RNA approach (Urich et al., 2008), metatranscriptomics analysis of total rRNA (Turner et al., 2013), total RNA sequencing (total RNA-Seq; Bang-Andreasen et al., 2020; Li & Guan, 2017; Li et al., 2016), total RNA-based metatranscriptomics or total RNA-seq-based metatranscriptomics (Li & Guan, 2017).

To distinguish this approach from regular metatranscriptomics, we will use the term *total RNA-Seq* in the balance of this paper. As 80–98% of cellular RNA consists of rRNA (Peano et al., 2013; Westermann et al., 2012), SSU (and LSU) rRNA can make up 37–71% of all RNA sequences (Elekwachi et al., 2017; Yu & Zhang, 2012). This means that a large portion of total RNA-Seq data can be used for taxonomic identification and full-length assembly of SSU rRNA.

Several studies took total RNA-Seq a step further and combined it with synthetic long-read sequencing (Dueholm et al., 2020; Karst et al., 2018; Tedersoo et al., 2021). Synthetic long-read sequencing is a modification of short-read sequencing in which each molecule is tagged with unique molecular identifiers (UMIs), amplified, randomly split and sequenced with conventional short-read sequencing technology. Fragments of the same molecule are linked through the tags, and complete molecules can be reconstructed from linked short sequences through de novo assembly, hence the term synthetic long reads. Karst et al. (2018) size-selected total RNA for SSU rRNA and applied synthetic long-read sequencing to generate over a million SSU rRNA sequences, of which almost 45,000 were complete. Although this approach is effective, it adds additional costs, time and room for bias. Due to the growing interest in applying multiple -omics approaches to environmental samples to generate a more holistic picture of ecosystems (Cordier et al., 2019, 2021; Leese et al., 2018; Uyaguari-Diaz et al., 2016), less specialized methods might be easier to implement, especially for routine application. Total RNA-Seq without size selection and UMI tagging follows the same protocol as conventional metatranscriptomics after mRNA enrichment, and both methods could be implemented simultaneously without additional modifications.

A few studies compared the performance of total RNA-Seq and metabarcoding (Lanzén et al., 2011; Yan et al., 2018) or metagenomics (Hempel et al., 2022; Lanzén et al., 2011; Shi et al., 2011; Urich et al., 2014; Uyaguari-Diaz et al., 2016) for the analysis of microbial community composition. However, the use of total RNA-Seq to analyse microbial communities remains rare, and a comparison of total RNA-Seq and metagenomics in terms of SSU rRNA reconstruction is lacking, although the results of such a comparison could impact our ability to categorize global biodiversity and analyse microbial communities.

In an earlier study, we were able to show that total RNA-Seq characterized the abundance profile of a microbial mock community consisting of 10 species more accurately than metagenomics at almost one order of magnitude lower sequencing depth (Hempel et al., 2022). For the present study, we used the same mock community data to evaluate the performance of metagenomics and total RNA-Seq in terms of SSU rRNA completeness, that is, the portion of the SSU rRNA that can be successfully reconstructed. Furthermore, we determined and compared the genome completeness of both approaches to highlight the advantages and disadvantages of either approach. Additionally, since total RNA-Seq has not yet been extensively tested, we also determined the impact of commonly used data processing tools on SSU rRNA reconstruction, as it has been repeatedly shown that results based on HTS are heavily influenced by the choice of bioinformatics tools (Bashiardes et al., 2016; Knight et al., 2018; McIntyre et al., 2017; Quince et al., 2017; Shakya et al., 2019; Vollmers et al., 2017). Lastly, we applied metagenomics and total RNA-Seq and the same data processing tools to an aquarium sample, which served as a proxy for an environmental freshwater sample. We evaluated SSU rRNA completeness for the aquarium community whose composition was unknown. This served as a proof of concept to determine if the conclusions based on the mock community aligned with those based on the simulated environmental sample.

2 | MATERIALS AND METHODS

The overall study design is shown in Figure 1, and further details are given in the balance of this section.

2.1 | Sampling

For the microbial mock community, we used a commercially available mock community (ZymoBIOMICS Microbial Community Standard II (Log Distribution); Zymo Research), consisting of eight bacterial species (three Gram-negative and five Gram-positive) and two yeast species with log-distributed species abundances based on genomic DNA quantity (Figure 2). The mock community was preserved in DNA/RNA Shield (Zymo Research) by the manufacturer to inactivate cells while preserving DNA and RNA. We generated three simulated water sample replicates by adding 130 µL of the mock community containing approximately 381 ng of total DNA to 50-mL ultrapure water respectively.

Furthermore, we took a 10-L water sample from an aquarium (Hagen Aqualab; University of Guelph; Figure S1) using a bleach-sterilized jug to simulate environmental freshwater sampling. The aquarium contained multiple fish, mollusc, crustacean and macrophyte species as well as an established microbial community. We mixed the 10-L sample and subsampled three 1-L samples.

2.2 | Laboratory and bioinformatics processing

The mock community data used in this study originate from an earlier study, in which we applied total RNA-Seq and metagenomics and investigated 672 combinations of data processing tools to identify the best-performing sequencing method and combination of tools to characterize the abundance profile of a microbial mock community (Hempel et al., 2022). Details of the laboratory and bioinformatics processing steps can be found there (Hempel et al., 2022). The three aquarium sample replicates were processed simultaneously with the three mock community replicates. In summary, we filtered samples through 0.2-µm Nalgene Analytical Test Filter Funnels (Thermo Fisher Scientific) and co-extracted DNA and RNA in parallel from each sample using a modified version of the Quick-DNA/

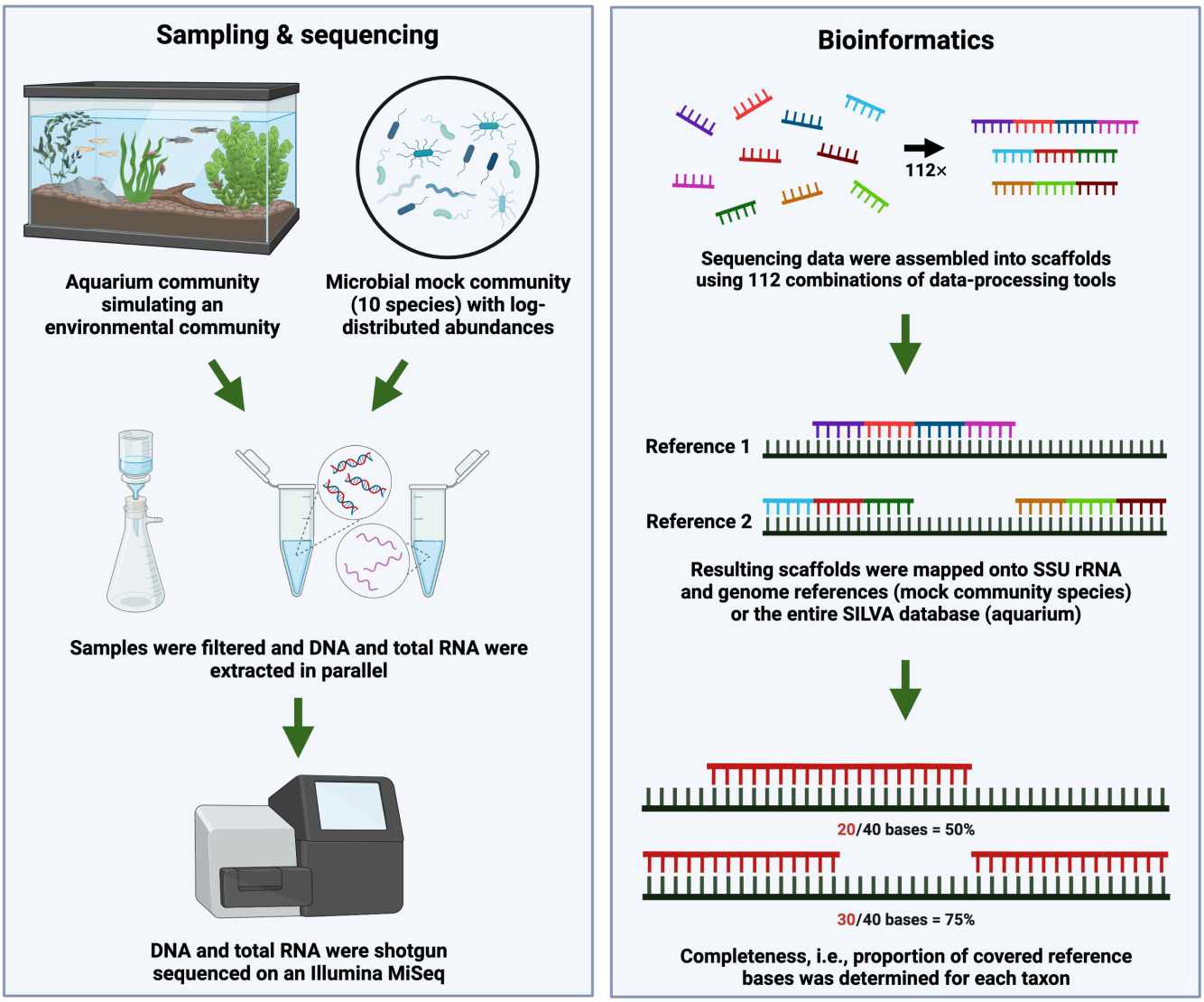

FIGURE 1 Summary of the study design (created using BioRender.com). The DNA and total RNA of a microbial mock community and an aquarium sample were extracted and shotgun-sequenced on an Illumina MiSeq, representing two sequencing methods (metagenomics and total RNA-Seq). Sequencing data were assembled into scaffolds using 112 combinations of common data processing tools. Scaffolds obtained from the mock community were mapped onto SSU rRNA and genome references of the 10 mock community species, and scaffolds obtained from the aquarium sample were mapped onto the entire SILVA database. The completeness, that is, proportion of covered bases, was determined for each matched taxon. rRNA, ribosomal RNA; SSU, small subunit.

RNA Microprep Plus Kit (Zymo Research), resulting in three replicate DNA and RNA extracts for both mock community and aquarium samples respectively. DNA and RNA library prep and sequencing were performed by a service provider (Génome Québec), and mRNA enrichment was skipped prior to RNA library prep to sequence total RNA. During library prep, normalization was performed by processing equal volumes of samples instead of equal concentrations of samples so that the relative number of reads per sample mirrored the relative amount of DNA/RNA, avoiding an over- or under-representation of samples with higher or lower amounts of DNA/RNA. The libraries were 150 bp PE sequenced on an Illumina MiSeq (2,428,038 PE reads in total; Bioproject number: PRJNA819997; for the number of reads per sample see Figure S1). Metagenomics mock community replicates received a much higher number of reads than total RNA-Seq mock community replicates, and to allow for appropriate comparisons between both approaches, we subsampled the reads of metagenomics replicates to match the number of reads of total RNA-Seq replicates. As DNA and RNA were coextracted, we subsampled each metagenomics mock community replicate to the number of reads of the corresponding total RNA-Seq replicate. Random subsampling was performed 10 times, and the subsamples were processed independently.

Sequence processing for this study was divided into three data processing steps: trimming and quality filtering, rRNA sorting and assembly. For trimming and quality filtering, four PHRED score cutoffs were applied (PHRED ≤5, ≤10, ≤15, and ≤20) to trim the end of reads using Trimmomatic v0.39 (Bolger et al., 2014). For rRNA sorting, three approaches were applied to sort reads into rRNA

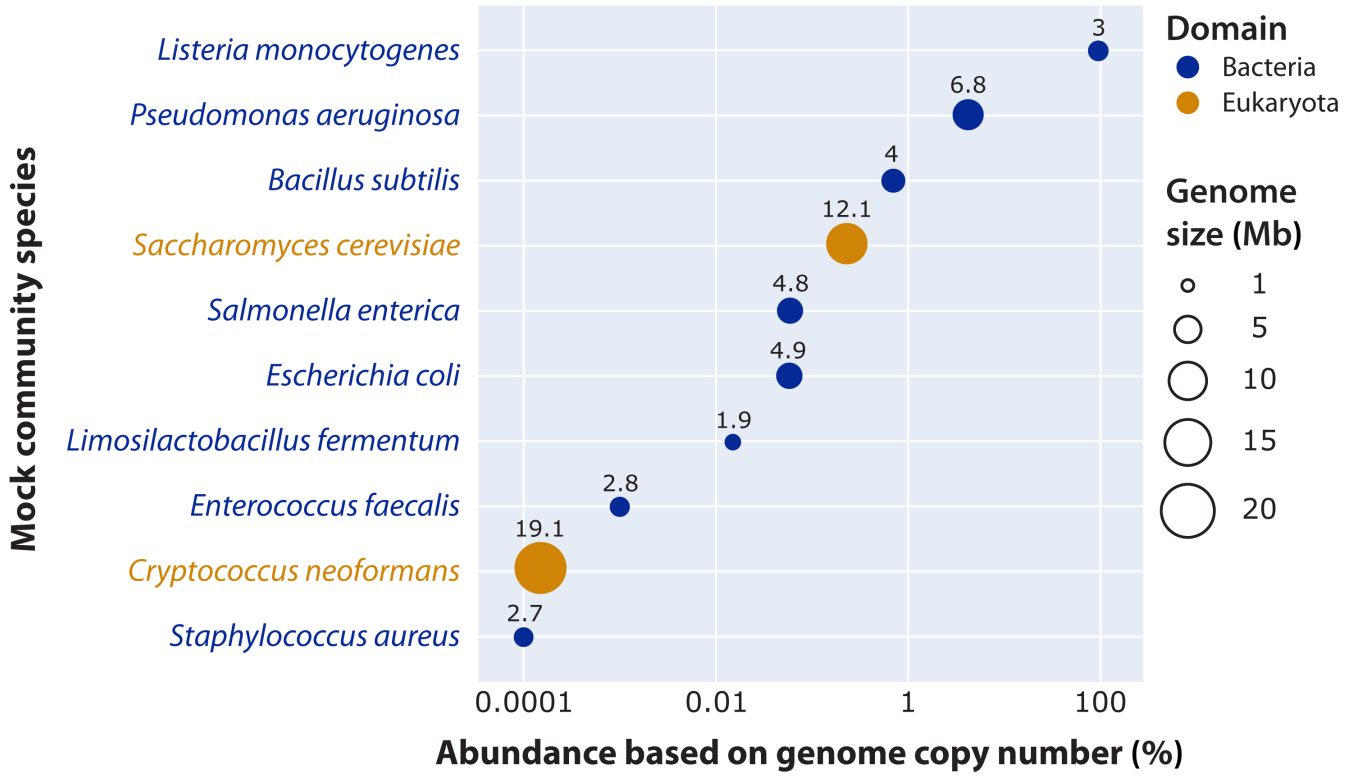

FIGURE 2 Genome size and relative abundance of mock community species. Relative abundances are based on genome copy numbers, as recommended by the manufacturer for studies that involve shotgun sequencing. The size of and numbers above bubbles indicate genome sizes in megabases (Mb).

and non-rRNA reads: alignment-based with SortMeRNA v4.0.0 (Kopylova et al., 2012), Hidden Markov model-based with barrnap v0.9 (Seemann, unpublished, https://github.com/tseemann/barrnap, accessed on 18 Jun 2021) and kmer-based with rRNAFilter v1.1 (Wang et al., 2017). For each approach, non-rRNA reads were subsequently excluded. Additionally, we omitted rRNA sorting and used all reads instead, leading to four rRNA sorting approaches in total. For assembly, we tested seven assemblers for both metagenomics and total RNA-Seq reads: SPAdes, metaSPAdes v3.14.1 (Nurk et al., 2017), MEGAHIT v1.2.9 (Li et al., 2015), IDBA-UD v1.1.1 (Peng et al., 2012), rnaSPAdes v3.14.1 (Bushmanova et al., 2019), IDBA-tran v1.1.1 (Peng et al., 2013) and Trans-ABySS v2.0.1 (Robertson et al., 2010). Trinity (Grabherr et al., 2013) is another widely used, metatranscriptomics-optimized assembler, but despite thorough efforts to run Trinity with adjusted RAM settings as recommended by the developers (https://trinityrnaseq.github.io/performance/mem.html), we ultimately failed to operate it consistently. Furthermore, there are multiple assemblers specifically designed for SSU rRNA assembly from metagenomics and even total RNA-Seq data, such as EMIRGE (Miller et al., 2011), MATAM (Pericard et al., 2018), MetaRib (Xue et al., 2020), RAMBL (Zeng et al., 2017) and REAGO (Yuan et al., 2015), but although we tried to run the two most recent of them, MATAM and MetaRib, MATAM runs consistently failed and we were not able to successfully install MetaRib (more specifically EMIRGE, which is a requirement for MetaRib). Therefore, we could not include any of them in our benchmarking, and the lack of user-friendliness or flexibility of these assemblers should be taken into account for future studies.

Combining all data processing tools in the three steps resulted in 112 combinations of tools that were applied to both metagenomics and total RNA-Seq data obtained from the mock community and aquarium samples. The code to run all combinations is available on GitHub (https://github.com/hempelc/metagenomics-vs-total RNASeq-reference-comparison).

2.3 | Determining SSU rRNA and genome completeness of mock community species

For each mock community species, we determined both SSU rRNA and genome completeness of the scaffolds generated in each of the 112 combinations separately. To determine completeness, references were required for each mock community species. Zymo Research provides full-length SSU rRNA and genome references for all mock community species, but SSU rRNA references of some species include multiple sequences to cover strain variants, and the genome references for the two eukaryotic species (*Saccharomyces cerevisiae* and *Cryptococcus neoformans*) consisted of draft genomes instead of curated genomes. Therefore, we aligned all SSU rRNA references of species with multiple reference sequences, respectively, using Geneious Prime v2022.1.1 (https://www.geneious.com) and the mafft plugin v7.450 (Katoh & Standley, 2013) with default parameters, extracted the consensus

sequence at a 100% threshold to include all ambiguities and used the consensus sequences as SSU rRNA references. We also downloaded the reference genomes of the eukaryotic species from RefSeq (GCF_000146045.2 for *S. cerevisiae* and GCF_000091045.1 for *C. neoformans*) and used those as reference genomes instead of the draft genomes provided by Zymo Research.

Both SSU rRNA and genome references for the 10 mock microbial community species were indexed for mapping using the *index* function of BWA v0.7.17 (Li & Durbin, 2009) with default parameters. The scaffolds resulting from each combination of tools were mapped onto both the indexed SSU rRNA and the genome reference of each species using the BWA-MEM algorithm of BWA with default parameters. We determined the completeness of each reference using the *coverage* function of samtools v1.10 (Li et al., 2009). Specifically, as the eukaryotic reference genomes consist of multiple chromosomes, we determined the total number of covered bases and divided it by the total number of bases to generate the relative completeness of each reference. This was done separately for each replicate, species, SSU rRNA and genome reference and combination of tools. Since metagenomics replicates were randomly subsampled 10 times, we determined the completeness of the metagenomics replicates by taking the mean average across the 10 subsamples. We summarized the completeness across metagenomics and total RNA-Seq replicates, respectively, by taking the mean average across the three replicates. The references used and the code to determine completeness are available on GitHub (https://github.com/hempelc/metagenomics-vs-totalRNASeq-reference-comparison).

2.4 | Determining the total number, SSU rRNA completeness and taxonomy of detected taxa in the aquarium samples

To compare the impact of sequencing types (metagenomics and total RNA-Seq) and data processing tools on the SSU rRNA completeness and taxonomy of environmental communities, we determined the total number, SSU rRNA completeness and taxonomy of detected taxa in the aquarium sample replicates. Therefore, we downloaded the SILVA SSU rRNA database SILVA132_NR99 (Quast et al., 2013), mapped all assembled scaffolds onto the database and calculated the completeness of each matched reference following the same method as described for the mock community. Additionally, we extracted the taxonomy of each matched taxon in the SILVA database and determined the number of detected prokaryotic and eukaryotic taxa. We only analysed the taxonomy at the domain level as the SILVA taxonomy is not standardized between prokaryotes and eukaryotes at higher taxonomic levels.

2.5 | Statistical analysis and visualization

All further data processing, statistical analysis and visualization were performed using Python v3.7.9 (Van Rossum & Drake, 2009). The full code is available on GitHub (https://github.com/hempelc/metagenomics-vs-totalRNASeq-reference-comparison) and involves the Python modules Pandas v1.3.5 (McKinney, 2010), NumPy v1.21.3 (Harris et al., 2020) and Plotly v5.6.0 (Plotly Technologies Inc, 2015).

2.5.1 | Mock community analysis

We sorted SSU rRNA and genome completeness across species by data processing tools and visualized differences in completeness using heatmaps. We visually observed a correlation between genome size and metagenomics SSU rRNA completeness, so we plotted the mean average SSU rRNA completeness across data processing tools for both metagenomics and total RNA-Seq against genome size and performed an Ordinary Least Squares regression to test for statistically significant correlations using the scatter function of Plotly with the parameter trendline set to 'ols'. Additionally, we plotted the mean average SSU rRNA completeness against abundance and performed an Ordinary Least Squares regression in the same way to test if species abundance correlated with SSU rRNA completeness.

As the genome completeness of all but one species was close to 0%, we only determined the impact of data processing tools on SSU rRNA completeness. For each species, we determined the correlation between each tool and SSU rRNA completeness by performing multiple linear regression with SSU rRNA completeness as the dependent variable and tools as binary independent variables using the OLS function of the Python module statsmodels v0.13.2 (Seabold & Perktold, 2010) and extracted coefficients and *p*-values. As the distribution of SSU rRNA completeness was non-normal for each species, we transformed completeness values into normal distribution using the QuantileTransformer function of the python module scikit-learn v1.0.2 (Pedregosa et al., 2011) for each species separately prior to multiple linear regression.

2.5.2 | Aquarium samples

We calculated and plotted the total number of detected taxa against the number of taxa with an SSU rRNA completeness of ≥80% for each aquarium sample replicate, sequencing type (metagenomics and total RNA-Seq) and combination of data processing tools. Since the number and SSU rRNA completeness of detected taxa were substantially impacted by the applied assemblers, we provided information on the assembler used for each combination in our plots. Furthermore, we calculated the ratio of the number of detected prokaryotic and eukaryotic taxa (prokaryote:eukaryote ratio) by dividing the number of detected prokaryotic taxa by the number of detected eukaryotic taxa for each aquarium sample replicate, sequencing type and combination of data processing tools. For each aquarium sample replicate and sequencing type, we visualized the prokaryote:eukaryote ratio across all combinations of data processing tools using boxplots with the box function of Plotly. In one metagenomics replicate, four combinations

of data processing tools did not detect any eukaryotes and their prokaryote:eukaryote ratio could not be determined, and therefore, they were excluded from this part of the analysis.

3 | RESULTS

3.1 | Mock community

We found considerable differences between metagenomics and total RNA-Seq in terms of SSU rRNA completeness of mock community species (Figure 3). For metagenomics, SSU rRNA completeness was very low for both eukaryotic species, *S. cerevisiae* (mean average: 40.69%) and *C. neoformans* (mean average: 36.32%), and low for three bacterial species (mean averages: *P. aeruginosa*—63.97%, *S. enterica*—67.5%, *E. coli*—58.99%) independent from the data processing tools used. In contrast, for total RNA-Seq, SSU rRNA sequences were near-complete for all but three species across all data processing tools except for the assemblers metaSPAdes and IDBA-tran (mean average for all but three species excluding the assemblers metaSPAdes and IDBA-tran: ≥94.74%).

The three species with the lowest SSU rRNA completeness (mean averages excluding the assemblers metaSPAdes and IDBA-tran: *L. fermentum*—85.97%, *E. faecalis*—88%,

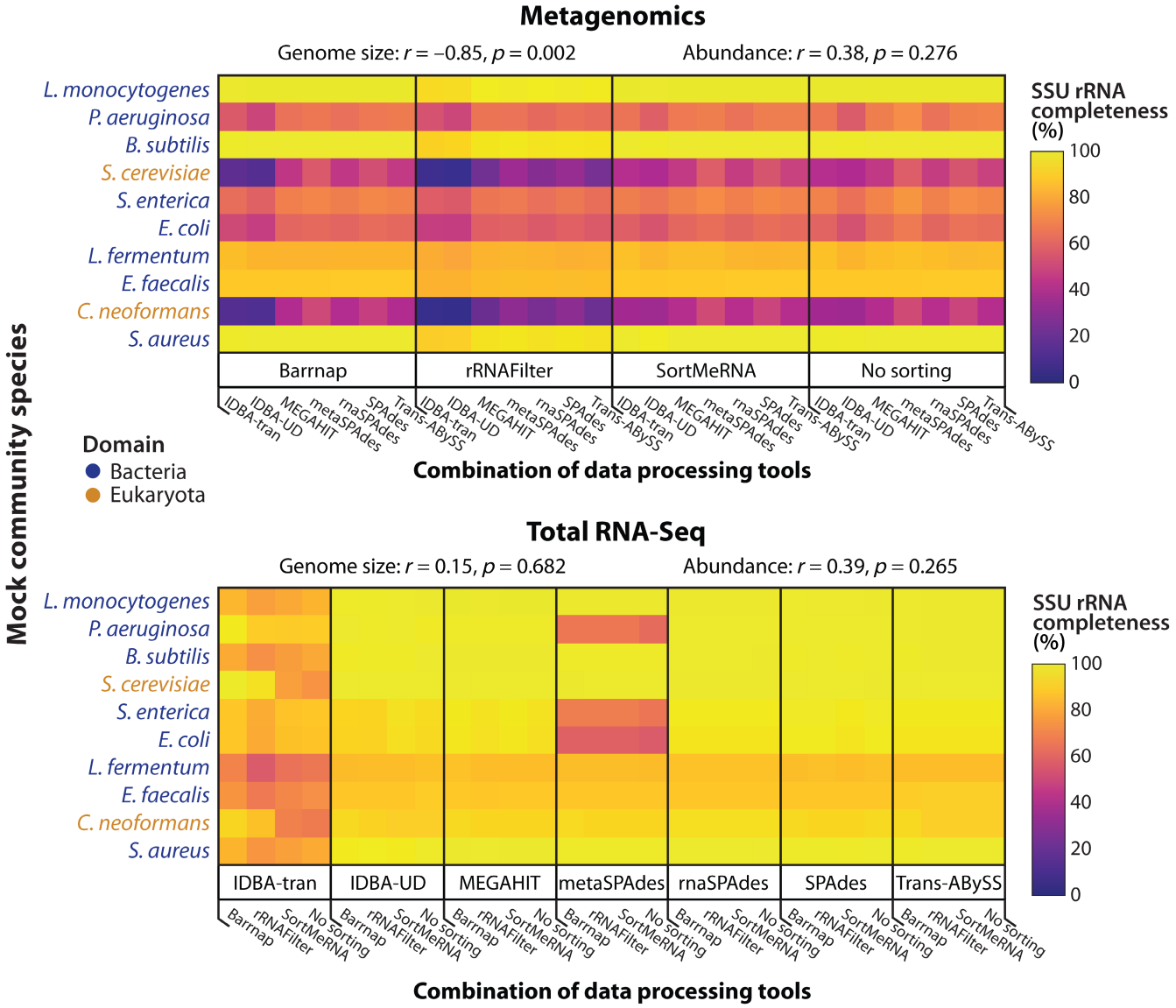

FIGURE 3 SSU rRNA completeness across mock community species and data processing combinations. Since PHRED scores used for trimming and quality filtering had mostly no significant impact on SSU rRNA completeness (Figure 4), only results based on PHRED score 5 are shown. The *x*-axis of each heatmap represents the combinations of data processing tools employed, the *y*-axis represents the 10 mock community species and the colour bar indicates SSU rRNA completeness in percentage. The combinations are sorted differently for each sequencing method to highlight the tool variations specific to the sequencing methods. Using total RNA-Seq, most species had near full-length sequences recovered (indicated by bright yellow) regardless of the combination of tools, except for the assemblers IDBA-tran and metaSPAdes. rRNA, ribosomal RNA; SSU, small subunit.

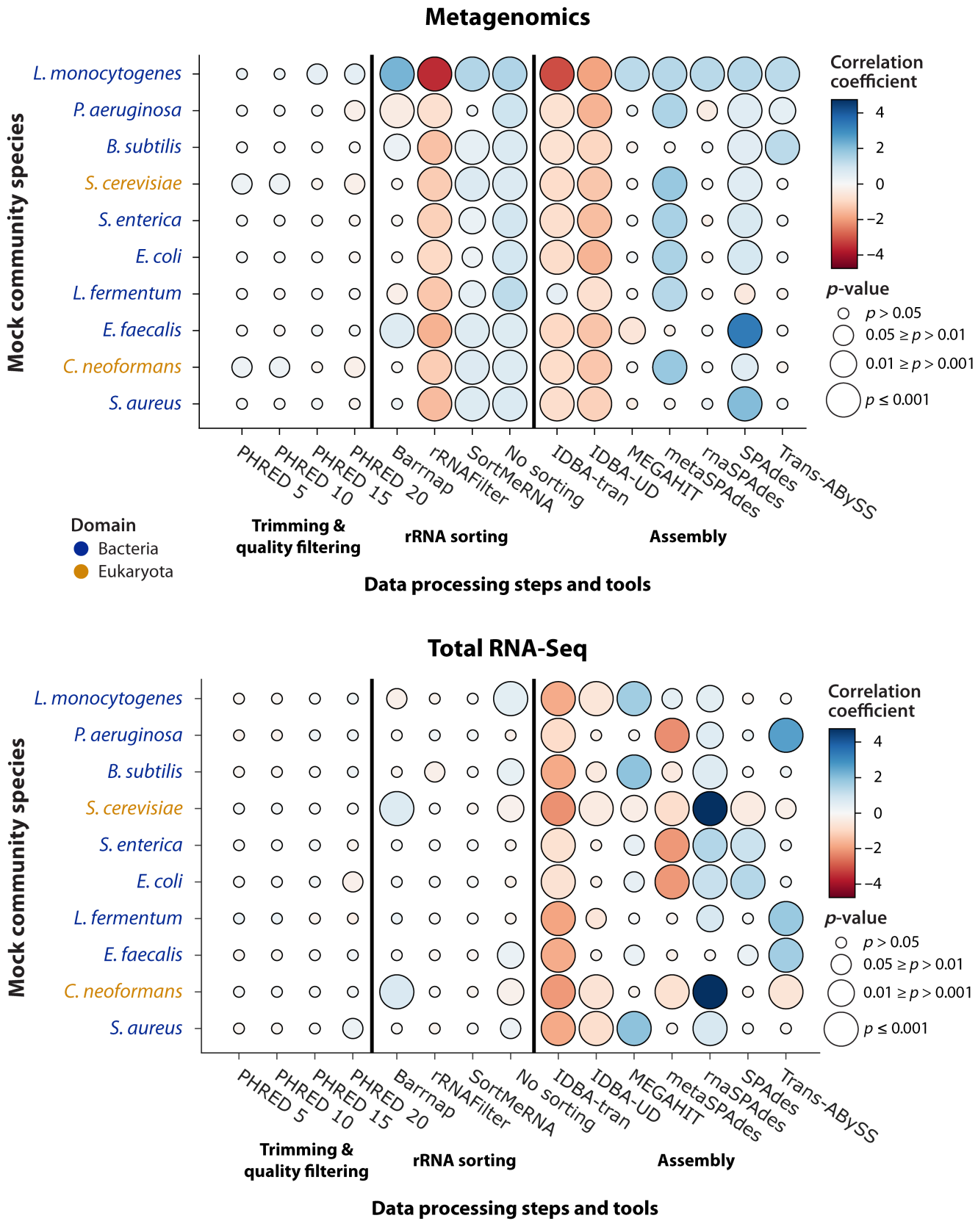

FIGURE 4 Impact of individual data processing tools on SSU rRNA completeness of the mock community. The *x*-axis of each bubble plot represents the tools employed, the *y*-axis represents the 10 mock community species, the colour bar indicates the correlation coefficient between the SSU rRNA completeness and the tools based on multiple linear regression and the bubble size indicates the significance of the correlations. Large-sized, red or blue bubbles indicate tools that are significantly negatively or positively correlated to SSU rRNA completeness. rRNA, ribosomal RNA; SSU, small subunit.

C. neoformans—91.52%) were among the four species with the lowest abundance.

The genome completeness of all species was close to 0% for either sequencing method across all utilized data processing tools and species except for *L. monocytogenes*, whose genome was almost 100% complete when applying metagenomics and no rRNA sorting and partially complete when applying metagenomics and rRNAFilter for rRNA sorting (Figure S2).

In terms of the significance and strength of correlations between each tool and SSU rRNA completeness, tools affected completeness differently for metagenomics and total RNA-Seq (Figure 4). For both sequencing methods, the tested PHRED scores used for trimming and quality filtering had no significant effect on the completeness, with only a few exceptions.

For metagenomics, the rRNA sorting tool rRNAFilter decreased SSU rRNA completeness, while SortMeRNA and no sorting increased SSU rRNA completeness (Figure 4, top). For total RNA-Seq, the impact of rRNA sorting tools was smaller, but no sorting increased the SSU rRNA completeness of four out of 10 species (Figure 4, bottom). The two eukaryotic species were the only species for which no sorting decreased completeness and Barrnap increased completeness for total RNA-Seq.

In terms of assembly tools, IDBA-tran decreased SSU rRNA completeness across all species for both sequencing methods (with one exception), and IDBA-UD decreased completeness across all species for metagenomics and half the species for total RNA-Seq. SPAdes and metaSPAdes performed well for metagenomics, while metaSPAdes performed poorly for total RNA-Seq. rnaSPAdes performed well for total RNA-Seq, in particular for the two eukaryotic species.

For metagenomics, SSU rRNA completeness was significantly negatively correlated to genome size but not correlated to species abundance (Figure 5, top). In contrast, SSU rRNA completeness was not correlated to either genome size or species abundance for total RNA-Seq (Figure 5, bottom). Neither sequencing method showed a correlation between SSU rRNA completeness and the interaction between genome size and species abundance (Figure S3).

3.2 | Aquarium samples

We found substantial differences between metagenomics and total RNA-Seq in terms of the total number and SSU rRNA completeness of taxa detected in the aquarium sample (Figure 6). Regardless of the utilized data processing tools, metagenomics detected almost no taxa based on mapping to the SSU rRNA database SILVA, and accordingly, almost no taxa with ≥80% SSU rRNA completeness. In contrast, total RNA-Seq detected up to 8636 taxa, of which up to 438 had an SSU rRNA completeness of ≥80%. Additionally, for total RNA-Seq, the results clustered based on the utilized assemblers, with only a few exceptions, demonstrating that among the three data processing steps, only assemblers strongly impacted the number of detected taxa and their SSU rRNA completeness. In particular, across all three technical replicates, MEGAHIT and rnaSPAdes had a positive impact on the SSU rRNA completeness of detected taxa, while Trans-ABySS and metaSPAdes had a positive impact on the number of detected taxa. IDBA-UD and SPAdes offered a trade-off between both, and their results were also impacted by the data processing tools utilized in the non-assembly steps, as indicated by their spread in the plot. IDBA-tran performed poorly in comparison with other assemblers.

In terms of the taxonomic composition of detected taxa, the prokaryote:eukaryote ratio indicates if the taxonomic composition on the domain level differed across utilized sequencing types and combinations of data processing tools. The median ratio across all combinations of data processing tools ranged from 3.43 to 4.84 across the metagenomics replicates and from 2.37 to 2.57 across the total RNA-Seq replicates (Figure 7). This shows that the taxonomic composition on the domain level differed between metagenomics and total RNA-Seq, with metagenomics detecting more prokaryotes than eukaryotes. Furthermore, the spread of the prokaryote:eukaryote ratio across combinations of data processing tools was much higher in metagenomics (ranges: 11.64, 32.3 and 13.84) than in total RNA-Seq (ranges: 3.73, 3.08 and 2.45). This indicates that data processing tools had a much higher impact on the taxonomic composition of metagenomics samples than on that of total RNA-Seq samples, and the relative taxonomic composition on the domain level remained relatively constant for total RNA-Seq regardless of the utilized data processing tools.

4 | DISCUSSION

4.1 | Microbial mock community

For the microbial mock community consisting of 10 species, total RNA-Seq generated SSU rRNA sequences with completeness equal to or higher than that of metagenomics. This was particularly the case for the two eukaryotic species with genomes much larger than those of the other mock community species, but also for bacterial species with large genomes. Furthermore, metagenomics SSU rRNA completeness was significantly negatively correlated with genome size, while this was not the case for total RNA-Seq. This shows that the broad genomic coverage of metagenomics leads to lower coverage of specific genes of interest, such as the SSU rRNA gene, especially with increasing genome size. The SSU rRNA coverage of total RNA-Seq was independent of genome size, confirming that total RNA-Seq naturally enriches the dataset for rRNA (Bang-Andreasen et al., 2020; Geisen et al., 2015; Urich et al., 2008).

Three of the four species with the lowest abundance (relative abundance = 0.015%–0.0001%) represented the species with the lowest SSU rRNA completeness when applying total RNA-Seq, likely due to their extremely low abundance and, therefore, relatively lower sequencing coverage. However, their completeness was on par for total RNA-Seq and metagenomics, and the SSU rRNA of the least abundant species was near-complete for either approach.

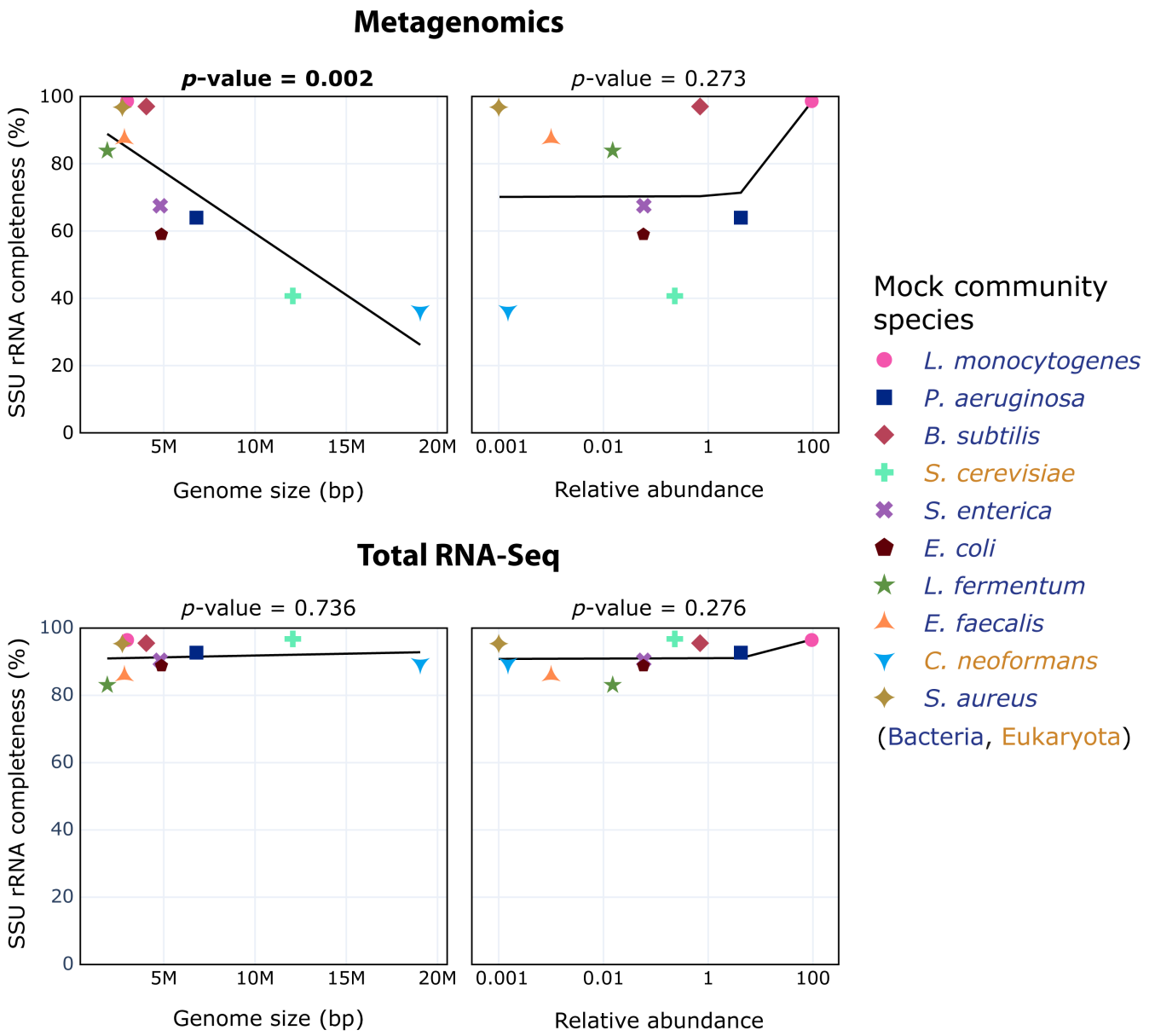

FIGURE 5 Relationship between SSU rRNA completeness and genome size (left)/relative species abundance (right) of the mock community. Trendlines are ordinary least squares regression lines, and all trendlines are linear but appear skewed for relative abundance plots due to the logarithmic *x*-axis. *p*-values indicate the significance of the correlations between SSU rRNA completeness and genome size/relative species abundance. rRNA, ribosomal RNA; SSU, small subunit.

SSU rRNA completeness of neither metagenomics nor total RNA-Seq significantly correlated with species abundance; nevertheless, given that three out of the four species with the lowest abundance were the only species without near-complete SSU rRNA sequences in total RNA-Seq, it might be possible that the inclusion of more low-abundant species would have revealed a statistically significant pattern, which should be further investigated. Against our expectation that SSU rRNA completeness of total RNA-Seq data is higher for low-abundant species, metagenomics recovered a sufficient amount of SSU rRNA fragments for SSU rRNA reconstruction of low-abundant species. An explanation might lie in the low complexity of the sample. The mock community used in our study contains only 10 species. Any difference between metagenomics and total RNA-Seq in terms of SSU rRNA completeness of low-abundant species might become only more prevalent in more complex communities.

The lack of a correlation between abundance and SSU rRNA completeness was in stark contrast to the observed correlation between abundance and genome completeness for metagenomics since the genome of the most abundant species was near-complete while the completeness of all other species' genomes was close to 0%. This illustrates that the abundance of each species in a community has a significant impact on its genome coverage when applying metagenomics, which is in agreement with other studies showing that successful genome reconstruction of low-abundance species

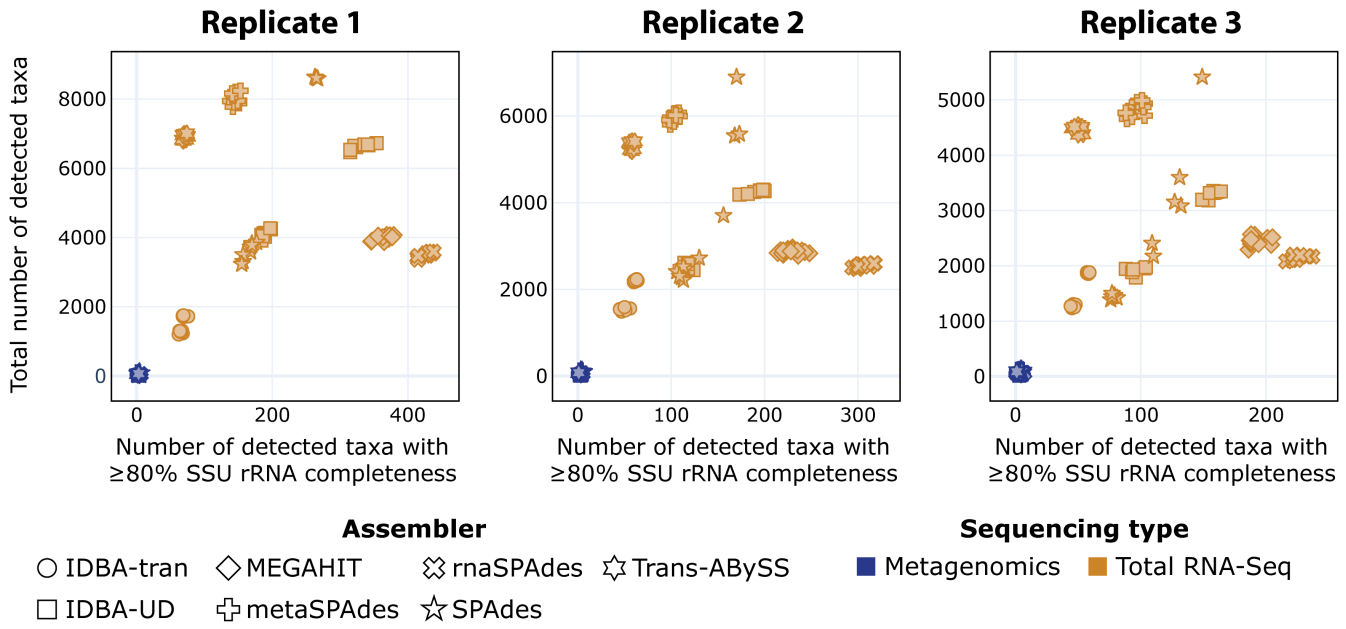

FIGURE 6 Total number of detected taxa and number of detected taxa with ≥80% SSU rRNA completeness across the three aquarium sample replicates, the two sequencing types (metagenomics and total RNA-Seq) and all utilized combinations of data processing tools. Symbols indicate the assemblers utilized for each combination of data processing tools. rRNA, ribosomal RNA; SSU, small subunit.

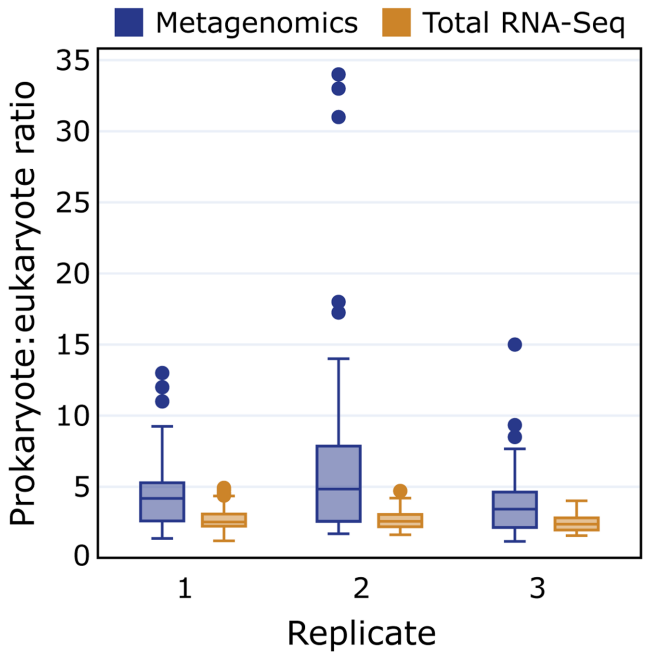

FIGURE 7 Prokaryote:eukaryote ratio across the three aquarium sample replicates, the two sequencing types (metagenomics and total RNA-Seq) and all utilized combinations of data processing tools.

typically requires high-sequencing depth (Jin et al., 2022; Merrill et al., 2022). In contrast to that, even the genome completeness of the most abundant species was close to 0% when applying total RNA-Seq, which further confirms that total RNA-Seq reads are naturally enriched for rRNA and consequently do not provide broad genomic coverage.

SSU rRNA completeness was independent of the applied PHRED score cut-off during quality filtering and trimming. It should be noted, however, that the quality of our sequences was almost exclusively over PHRED 30 according to mean and per sequence quality scores (Figure S4), so trimming and quality filtering might not have had a great effect in our study but could still have significant effects in studies where only lower quality data are available.

The methods applied for rRNA sorting impacted the SSU rRNA completeness of metagenomics but mostly not that of total RNA-Seq. In particular, rRNA filter, which sorts sequences based on k-mer abundances, was strongly negatively correlated with metagenomics SSU rRNA completeness, while SortMeRNA and no sorting significantly improved completeness. This shows that while extracting rRNA sequences from metagenomics sequences for SSU rRNA reconstruction can increase completeness when using appropriate tools, it can also decrease completeness when using unsuitable tools. Since no sorting significantly improved completeness, the application of rRNA sorting tools for metagenomics can be skipped, which also saves time and computational resources. rRNA sorting had mostly no effect on total RNA-Seq SSU rRNA completeness, likely because the sequences were already naturally enriched for rRNA. However, for the two eukaryotic species, no sorting significantly decreased completeness, while the tool barrnap significantly increased completeness. We conclude that rRNA sorting can also be omitted for total-RNA-Seq-based SSU rRNA reconstruction from bacteria, but further research on the impact of rRNA sorting on eukaryotic SSU rRNA reconstruction success is required.

In terms of genome completeness, when metagenomics was applied, only unsorted sequences allowed for complete genome reconstruction of the most abundant species, while genome completeness

was close to zero when SortMeRNA or barrnap was applied. This confirms that SortMeRNA and barrnap successfully sort the sequences into rRNA and non-rRNA. Surprisingly, the genome completeness was still comparably high at about 50% on average when applying rRNAFilter, which should not have been the case and indicates that rRNAFilter classified large amounts of non-rRNA sequences as rRNA. Given that rRNAFilter was also strongly negatively correlated with SSU rRNA completeness, we conclude that it should not be used for rRNA sorting in general. Only one other benchmarking study tested rRNAFilter for rRNA sorting and showed that the tool performs poorly in comparison to other rRNA sorting tools (Deng et al., 2022), which supports our conclusion.

The assemblers used in this study impacted the SSU rRNA completeness for both metagenomics and total RNA-Seq. Both IDBA-UD, which is optimized for metagenomics, and IDB-tran, which is optimized for metatranscriptomics, performed poorly for both sequencing methods. In contrast, metaSPAdes, which is optimized for metagenomics, performed well for metagenomics and poorly for total RNA-Seq, and rnaSPAdes, which is optimized for metatranscriptomics, performed well for total RNA-Seq, in particular for both eukaryotic species. These results show that specific assemblers can reduce or increase SSU rRNA completeness, and metaSPAdes seems the best when reconstructing SSU rRNA sequences from metagenomics data while rnaSPAdes seems ideal when reconstructing SSU rRNA sequences from total RNA-Seq data.

4.2 | Aquarium samples

For the aquarium samples, total RNA-Seq detected much more total taxa and taxa with ≥80% SSU rRNA completeness than metagenomics based on mapping against the SSU rRNA database SILVA. These results aligned with the results for the mock community and further support the hypothesis that total RNA-Seq naturally enriches sequences for SSU rRNA while metagenomics results in a much lower coverage of the SSU rRNA gene.

The number of detected taxa using total RNA-Seq was surprisingly high, with up to 8636 taxa detected in replicate 1. It is possible that the mapping of scaffolds onto the SILVA SSU rRNA database resulted in a high number of false-positive detections, and a more in-depth analysis of the taxonomic diversity using BLAST (Altschul et al., 1990) or kraken2 (Wood et al., 2019) in combination with a lowest common ancestor approach is required to confirm the high number of detected taxa. Nevertheless, SSU rRNA sequences with a completeness of ≥80% provide a reliable insight into the SSU rRNA coverage of metagenomics and total RNA-Seq, and while the former detected almost no taxa with an SSU rRNA completeness of ≥80%, the latter detected up to 438 such taxa in replicate 1 with around 1.9 M 100-bp-long paired-end reads (1/5 of an Illumina MiSeq run). Given that the SILVA database covers only a small fraction of the estimated microbial diversity on Earth, it is likely that a substantial number of microbial taxa present in the aquarium sample is not represented in the reference database. Therefore, the actual number of taxa sequenced with an SSU rRNA completeness of ≥80% might be even higher.

In a similar study, Karst et al. (2018) applied a more specialized approach, for which they size-selected SSU rRNA and applied synthetic long-read sequencing, and they detected around 45,000 bacterial taxa with an SSU rRNA completeness of around ≥80% across 19 samples, seven different environments, 13 Illumina MiSeq runs and 14 Illumina HiSeq runs. Our approach needs to be performed on a similar scale to allow for a direct comparison; nevertheless, our results are promising for further upscaling and demonstrate that total RNA-Seq without additional modifications also allows for the generation of high numbers of near-complete to complete SSU rRNA sequences from environmental samples, providing a method that could be implemented into routine environmental applications.

The utilized assemblers had a substantial impact on the SSU rRNA completeness and should be selected carefully in similar studies. To maximize SSU rRNA completeness, MEGAHIT and rnaSPAdes seem valuable options, which aligns with the results of the mock community analysis. Data processing tools also impacted the taxonomic composition of detected taxa on the domain level, although this impact was much higher for metagenomics than for total RNA-Seq, indicating that the taxonomic composition of total-RNA-Seq-based communities might be more robust to variations among data processing tools than that of metagenomics-based communities.

5 | CONCLUSIONS

In contrast to metagenomics, total RNA-Seq allowed for the complete or near-complete reconstruction of SSU rRNA sequences for all 10 species of a microbial mock community, and when analysing an aquarium sample that served as a proxy for an environmental sample, total RNA-Seq generated up to 438 SSU rRNA sequences with ≥80% completeness using only 1/5 of an Illumina MiSeq run, while metagenomics generated almost no such sequences. Furthermore, based on the microbial mock community, we show that total-RNA-Seq-based SSU rRNA completeness was independent of genome size, while metagenomics-based SSU rRNA completeness was significantly negatively correlated with the genome size of taxa in the community. Specific data processing tools impacted SSU rRNA completeness significantly, specifically the utilized assembler, and similar studies should select assemblers carefully. These results are promising for the high-throughput reconstruction of novel full-length SSU rRNA sequences and the simultaneous application of multiple -omics approaches to routine assessments of ecosystems, which is advocated by multiple studies (Cordier et al., 2019, 2021; Leese et al., 2018; Uyaguari-Diaz et al., 2016). While Karst et al. (2018) applied synthetic long-read sequencing to generate an unprecedented amount of novel SSU rRNA sequences from environmental samples, we demonstrate the effectiveness of a more conventional, short-read-based approach on a smaller scale. It needs to be explored which approach

proves more effective for large-scale applications in routine environmental assessments.

AUTHOR CONTRIBUTIONS

Christopher A. Hempel conceived the ideas and designed the methodology; Christopher A. Hempel collected the data; Christopher A. Hempel, Shea E. E. Carlson and Tyler A. Elliott analysed the data; Christopher A. Hempel and Shea E. E. Carlson led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Financial support for this work was provided by the Food from Thought: Agricultural Systems for a Healthy Planet Initiative, by the Canada First Research Excellence Fund (project 000054) through the University of Guelph; grants in Bioinformatics and Computational Biology from the Government of Canada through Genome Canada and Ontario Genomics and from the Ontario Ministry of Economic Development, Job Creation and Trade (project 15401); and from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN-2022-04569).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14149.

DATA AVAILABILITY STATEMENT

The sequencing data are available under Bioproject number PRJNA819997.

ORCID

Christopher A. Hempel https://orcid.org/0000-0002-2324-3115
Sarah J. Adamowicz https://orcid.org/0000-0001-9511-1229
Dirk Steinke https://orcid.org/0000-0002-8992-575X

REFERENCES

Almeida, O. G. G., & De Martinis, E. C. P. (2019). Bioinformatics tools to assess metagenomic data for applied microbiology. *Applied Microbiology and Biotechnology*, *103*(1), 69–82. https://doi.org/10.1007/s00253-018-9464-9

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.

Bang-Andreasen, T., Anwar, M. Z., Lanzén, A., Kjøller, R., Rønn, R., Ekelund, F., & Jacobsen, C. S. (2020). Total RNA sequencing reveals multilevel microbial community changes and functional responses to wood ash application in agricultural and forest soil. *FEMS Microbiology Ecology*, *96*(3), 1–13. https://doi.org/10.1093/femsec/fiaa016

Bashiardes, S., Zilberman-Schapira, G., & Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinformatics and Biology Insights*, *10*, 19–25. https://doi.org/10.4137/BBI.S34610

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, *15*(6), 1403–1414. https://doi.org/10.1111/1755-0998.12399

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), 1–13. https://doi.org/10.1093/gigascience/giz100

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*(13), 2937–2958. https://doi.org/10.1111/mec.15472

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing environmental genomics and machine learning for routine biomonitoring. *Trends in Microbiology*, *27*(5), 387–397. https://doi.org/10.1016/j.tim.2018.10.012

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS One*, *16*, e0257521. https://doi.org/10.1371/journal.pone.0257521

Deng, Z.-L., Münch, P. C., Mreches, R., & McHardy, A. C. (2022). Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Research*, *50*(10), e60. https://doi.org/10.1093/nar/gkac112

Dueholm, M. S., Andersen, K. S., McIlroy, S. J., Kristensen, J. M., Yashiro, E., Karst, S. M., Albertsen, M., & Nielsen, P. H. (2020). Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax). *MBio*, *11*(5), 1–14. https://doi.org/10.1128/mBio.01557-20

Elekwachi, C. O., Wang, Z., Wu, X., Rabee, A., & Forster, R. J. (2017). Total rRNA-Seq analysis gives insight into bacterial, fungal, protozoal and archaeal communities in the rumen using an optimized RNA isolation method. *Frontiers in Microbiology*, *8*(1814), 1–14. https://doi.org/10.3389/fmicb.2017.01814

Fan, L., McElroy, K., & Thomas, T. (2012). Reconstruction of ribosomal RNA genes from metagenomic data. *PLoS One*, *7*(6), 1–9. https://doi.org/10.1371/journal.pone.0039948

Geisen, S., Tveit, A. T., Clark, I. M., Richter, A., Svenning, M. M., Bonkowski, M., & Urich, T. (2015). Metatranscriptomic census of active protists in soils. *ISME Journal*, *9*(10), 2178–2190. https://doi.org/10.1038/ismej.2015.30

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2013). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883.Trinity

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M., Brett, M., Haldane, A., del Río, J., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hempel, C. A., Wright, N., Harvie, J., Hleap, J. S., Adamowicz, S. J., & Steinke, D. (2022). Metagenomics versus total RNA sequencing: Most accurate data-processing tools, microbial identification accuracy, and perspectives for freshwater assessments. *Nucleic*

Acids Research, *50*, 9279–9293. https://doi.org/10.1093/nar/gkac689

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, *7*(1), 1–11. https://doi.org/10.1038/s41597-020-00743-4

Jin, H., You, L., Zhao, F., Li, S., Ma, T., Kwok, L. Y., Xu, H., & Sun, Z. (2022). Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome. *Gut Microbes*, *14*(1), 1–18. https://doi.org/10.1080/19490976.2021.2021790

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-13036-1

Karst, S. M., Dueholm, M. S., McIlroy, S. J., Kirkegaard, R. H., Nielsen, P. H., & Albertsen, M. (2018). Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nature Biotechnology*, *36*(2), 190–195. https://doi.org/10.1038/nbt.4045

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Lanzén, A., Jørgensen, S. L., Bengtsson, M. M., Jonassen, I., Øvreås, L., & Urich, T. (2011). Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. *FEMS Microbiology Ecology*, *77*(3), 577–589. https://doi.org/10.1111/j.1574-6941.2011.01138.x

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., Ekrem, T., Čiampor, F., Čiamporová-Zaťovičová, Z., Costa, F., Duarte, S., Elbrecht, V., Fontaneto, D., Franc, A., Geiger, M., Hering, D., Kahlert, M., Stroil, B. K., Kelly, M., … Weigand, A. M. (2018). Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: A perspective derived from the DNAqua-net COST action. *Advances in Ecological Research*, *58*, 63–99. https://doi.org/10.1016/bs.aecr.2018.01.001

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, F., & Guan, L. L. (2017). Metatranscriptomic profiling reveals linkages between the active rumen microbiome and feed efficiency in beef cattle. *Applied and Environmental Microbiology*, *83*(9), 1–16. https://doi.org/10.1128/AEM.00061-17

Li, F., Henderson, G., Sun, X., Cox, F., Janssen, P. H., & Guan, L. L. (2016). Taxonomic assessment of rumen microbiota using total RNA and targeted amplicon sequencing approaches. *Frontiers in Microbiology*, *7*, 987. https://doi.org/10.3389/fmicb.2016.00987

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(21), 5970–5975. https://doi.org/10.1073/pnas.1521291113

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., Raes, J., Poulain, J., Jaillon, O., Wincker, P., Kandels-Lewis, S., Karsenti, E., Bork, P., & Acinas, S. G. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, *16*(9), 2659–2671. https://doi.org/10.1111/1462-2920.12250

McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L., & Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, *18*(1), 1–19. https://doi.org/10.1186/s13059-017-1299-7

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in science conference*, 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Merrill, B. D., Carter, M. M., Olm, M. R., Dahan, D., Tripathi, S., Spencer, S. P., Yu, B., Jain, S., Neff, N., Jha, A. R., & Sonnenburg, J. L. (2022). Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing microbes. *BioRxiv*. https://doi.org/10.1101/2022.03.30.486478

Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., & Banfield, J. F. (2011). EMIRGE: Reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*, *12*(5), R44. https://doi.org/10.1186/gb-2011-12-5-r44

Milo, R., & Phillips, R. (2015). *Cell biology by the numbers (draft)*. Garland Science, Taylor & Francis Group. http://book.bionumbers.org/what-are-the-rates-of-cytoskeleton-assembly-and-disassembly/

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I. M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., … de Vargas, C. (2012). CBOL Protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, *10*(11), e1001419. https://doi.org/10.1371/journal.pbio.1001419

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J. A., & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, *55*, 12–25. https://doi.org/10.1016/j.ejop.2016.02.003

Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., de Bellis, G., & Landini, P. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microbial Informatics and Experimentation*, *3*(1), 1–11. https://doi.org/10.1186/2042-5783-3-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M.,

& Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Peng, Y., Leung, H. C. M., Yiu, S. M., Lv, M. J., Zhu, X. G., & Chin, F. Y. L. (2013). IDBA-Tran: A more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, *29*(13), 326–334. https://doi.org/10.1093/bioinformatics/btt219

Pericard, P., Dufresne, Y., Couderc, L., Blanquart, S., & Touzet, H. (2018). MATAM: Reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. *Bioinformatics*, *34*(4), 585–591. https://doi.org/10.1093/bioinformatics/btx644

Plotly Technologies Inc. (2015). *Collaborative data science*. Plotly Technologies Inc. https://plot.ly

Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., III, Weinstock, G. M., White, O., & Huttenhower, C. (2019). The integrative human microbiome project. *Nature*, *569*(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. https://doi.org/10.1093/nar/gks1219

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844. https://doi.org/10.1038/nbt.3935

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

Sagova-Mareckova, M., Boenigk, J., Bouchez, A., Cermakova, K., Chonova, T., Cordier, T., Eisendle, U., Elersek, T., Fazi, S., Fleituch, T., Frühe, L., Gajdosova, M., Graupner, N., Haegerbaeumer, A., Kelly, A. M., Kopecky, J., Leese, F., Nõges, P., Orlic, S., … Stoeck, T. (2021). Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Research*, *191*, 116767. https://doi.org/10.1016/j.watres.2020.116767

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in science conference*, 57–61. https://doi.org/10.25080/majora-92bf1922-011

Shakya, M., Lo, C. C., & Chain, P. S. G. (2019). Advances and challenges in metatranscriptomic analysis. *Frontiers in Genetics*, *10*(SEP), 1–10. https://doi.org/10.3389/fgene.2019.00904

Shi, Y., Tyson, G. W., Eppley, J. M., & Delong, E. F. (2011). Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME Journal*, *5*(6), 999–1013. https://doi.org/10.1038/ismej.2010.189

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., Gies, E. A., Cheng, J. F., Copeland, A., Klenk, H. P., Hallam, S. J., Hugenholtz, P., Tringe, S. G., & Woyke, T. (2016). High-resolution phylogenetic microbial community profiling. *ISME Journal*, *10*(8), 2020–2032. https://doi.org/10.1038/ismej.2015.249

Smith, M. B., Rocha, A. M., Smillie, C. S., Olesen, S. W., Paradis, C., Wu, L., Campbell, J. H., Fortney, J. L., Mehlhorn, T. L., Lowe, K. A., Earles, J. E., Phillips, J., Techtmann, S. M., Joyner, D. C., Elias, D. A., Bailey, K. L., Hurt, R. A., Jr., Preheim, S. P., Sanders, M. C., … Hazen, T. C. (2015). Natural bacterial communities serve as quantitative geochemical biosensors. *MBio*, *6*(3), e00326-15. https://doi.org/10.1128/mBio.00326-15

Tedersoo, L., Albertsen, M., & Anslan, S. (2021). Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied Environmental Microbiology*, *87*(17), 1–19.

Turner, T. R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., Osbourn, A., Grant, A., & Poole, P. S. (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME Journal*, *7*(12), 2248–2258. https://doi.org/10.1038/ismej.2013.119

Urich, T., Lanzén, A., Qi, J., Huson, D. H., Schleper, C., & Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One*, *3*(6), e2527. https://doi.org/10.1371/journal.pone.0002527

Urich, T., Lanzén, A., Stokke, R., Pedersen, R. B., Bayer, C., Thorseth, I. H., Schleper, C., Steen, I. H., & Ovreas, L. (2014). Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environmental Microbiology*, *16*(9), 2699–2710. https://doi.org/10.1111/1462-2920.12283

Uyaguari-Diaz, M. I., Chan, M., Chaban, B. L., Croxen, M. A., Finke, J. F., Hill, J. E., Peabody, M. A., van Rossum, T., Suttle, C. A., Brinkman, F. S., Isaac-Renton, J., Prystajecky, N. A., & Tang, P. (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome*, *4*, 1–19. https://doi.org/10.1186/s40168-016-0166-1

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vollmers, J., Wiegand, S., & Kaster, A. K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—Not only size matters! *PLoS One*, *12*(1), 1–31. https://doi.org/10.1371/journal.pone.0169662

Wang, Y., Hu, H., & Li, X. (2017). rRNAFilter: A fast approach for ribosomal RNA read removal without a reference database. *Journal of Computational Biology*, *24*(4), 368–375. https://doi.org/10.1089/cmb.2016.0113

Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, *10*(9), 618–630. https://doi.org/10.1038/nrmicro2852

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

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/s13059-019-1891-0

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, *6*(2), e1000667. https://doi.org/10.1371/journal.pcbi.1000667

Xue, Y., Lanzén, A., & Jonassen, I. (2020). Reconstructing ribosomal genes from large scale total RNA meta-transcriptomic data. *Bioinformatics*, *36*(11), 3365–3371. https://doi.org/10.1093/bioinformatics/btaa177

Yan, Y. W., Jiang, Q. Y., Wang, J. G., Zhu, T., Zou, B., Qiu, Q. F., & Quan, Z. X. (2018). Microbial communities and diversities in mudflat sediments analyzed using a modified metatranscriptomic method. *Frontiers in Microbiology*, *9*(93), 1–15. https://doi.org/10.3389/fmicb.2018.00093

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645. https://doi.org/10.1038/nrmicro3330

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), 415–420. https://doi.org/10.1038/nbt.1823

Yu, K., & Zhang, T. (2012). Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS One*, *7*(5), e38183. https://doi.org/10.1371/journal.pone.0038183

Yuan, C., Lei, J., Cole, J., & Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*, *31*(12), i35–i43. https://doi.org/10.1093/bioinformatics/btv231

Zeng, F., Wang, Z., Wang, Y., Zhou, J., & Chen, T. (2017). Large-scale 16S gene assembly using metagenomics shotgun sequences. *Bioinformatics*, *33*(10), 1447–1456. https://doi.org/10.1093/bioinformatics/btx018

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Number of reads per sample.

Figure S2. Genome completeness across mock community species and data-processing combinations.

Figure S3. Relationship between SSU rRNA completeness and genome size × species abundance of the mock community.

Figure S4. Mean and per-sequence quality scores for metagenomics and total RNA-Seq data.

How to cite this article: Hempel, C. A., Carson, S. E. E., Elliott, T. A., Adamowicz, S. J., & Steinke, D. (2023). Reconstruction of small subunit ribosomal RNA from high-throughput sequencing data: A comparative study of metagenomics and total RNA sequencing. *Methods in Ecology and Evolution*, *00*, 1–16. https://doi.org/10.1111/2041-210X.14149