Comparing metagenomics, total RNA sequencing, and bioinformatics pipelines for multi-domain taxonomic profiling: implications for ecological assessments

# Abstract

Ecological assessments are necessary to evaluate the status of our deteriorating ecosystems, however, assessment methods traditionally omit most microbes since unicellular organisms are challenging to identify. This omission is not ideal, as microbes might better represent current environmental conditions than traditional taxa. DNA- and RNA-based techniques are increasingly applied for ecological assessments to overcome this identification challenge but require more testing and optimization. In this study, we compare metagenomics and total RNA sequencing (total RNA-Seq) in terms of their taxonomic profiling performance for microbial communities. We applied both techniques on two sample sets, 1) a commercially available microbial mock community, and 2) a display tank water sample. We processed the data using 1,532 bioinformatics pipelines and evaluated each workflow, i.e., the combination of sample type (metagenomics or total RNA-Seq) and pipeline, in terms of their accuracy and precision.

Results and discussion are to be added to abstract

Keywords: biomonitoring, freshwater, ecological assessment, metagenomics, metatranscriptomics, benchmarking, taxonomic profiling, multi-domain, bioinformatics pipeline, mock community

# Introduction

Our ecosystems are globally deteriorating at an unprecedented speed, causing a rapid biodiversity decline that is considered the sixth mass extinction event (Ceballos et al., 2015; IPBES, 2019; Pimm et al., 2014; WWF, 2020). This deterioration negatively affects ecosystem services, and the economic losses ﻿due to land use change are estimated at 4.3 – 20.2 trillion $US/year between 1997 to 2011 (Costanza et al., 2014). Consequently, ecosystem protection is gaining increased attention, also on a political scale (IPBES, 2019). In this study, we focus on the assessment of freshwater ecosystems, however, the implications of our study could be valuable for other ecosystem types as well, including soil and marine environments.

Freshwater ecosystems are valuable for economic productivity, ecosystem resilience, and for maintaining ecosystem services, which includes the supply of clean, consumable water (Dudgeon et al., 2006). However, freshwaters are heavily impacted by agricultural and industrial pollution, habitat fragmentation, and the introduction of invasive species (Dextrase and Mandrak, 2006; Jensen et al., 2006). As a result, freshwaters are among the most threatened ecosystems exhibiting some of the highest rates of species loss (Malmqvist and Rundle, 2002).

To prevent this species loss, the natural status of freshwater ecosystems needs to be protected, preserved, and restored. Therefore, we first need to determine the ecological status of freshwater bodies by inventorying their biodiversity. Such inventories can be screened for the presence and abundance of species that represent specific environmental conditions, so-called bioindicators (Burger, 2006). Common bioindicators are animals, plants, and microbial diatoms (Bellinger and Sigee, 2015; Haury et al., 2006; Karr, 1981; Resh and Unzicker, 1975); however, there are ongoing efforts to include other microbes (unicellular organisms, including bacteria, archaea, and unicellular eukaryotes), since they respond faster to environmental changes and therefore might better represent current environmental conditions (Cordier et al., 2019; Foissner and Berger, 1996; McArthur, 2001; Pawlowski et al., 2016; Payne, 2013; Smith et al., 2015; Stoeck et al., 2018).

Biodiversity inventories are generated by taxonomically identifying the community of organisms in a sample. This process is called taxonomic profiling and is traditionally done by assessing the morphology of organisms. However, morphological identification can be biased (Stein et al., 2014; Sweeney et al., 2011) or is not feasible due to a lack of diagnostical traits, especially when it comes to microbes (Pawlowski et al., 2012; Will and Rubinoff, 2004). DNA metabarcoding (Taberlet et al., 2012) was suggested as a complementary, DNA-based approach for taxonomic profiling of communities. However, it can be biased as well due to varying primer-binding affinities (Alberdi et al., 2018; Elbrecht and Leese, 2015; Krehenwinkel et al., 2017; Piñol et al., 2014; Piper et al., 2019) and PCR-related issues (Jane et al., 2015; Nichols et al., 2018; Piper et al., 2019; Sze and Schloss, 2019; Taberlet et al., 1996). Furthermore, DNA metabarcoding primers are designed to target specific taxonomic groups, and taxonomically diverse communities cannot be entirely covered.

Taxonomic coverage can be increased by including multiple primers (Alberdi et al., 2018; De Barba et al., 2014; Stat et al., 2017; Zhang et al., 2018), but this approach is also known to introduce additional bias (Corse et al., 2019; De Barba et al., 2014; Piñol et al., 2019). Consequently, both traditional approaches and DNA metabarcoding have limitations for taxonomic profiling of the biodiversity in a freshwater ecosystem.

An alternative approach for taxonomic profiling of communities is shotgun sequencing. It involves the random fragmentation and sequencing of the entire DNA in a sample, which is called metagenomics (Almeida and De Martinis, 2019; Wooley et al., 2010), or of the entire RNA in a sample, which is called metatranscriptomics (Shakya et al., 2019). That way, entire communities can be sequenced and identified primer- and PCR-free, thereby, removing the bias associated with traditional approaches and DNA metabarcoding. Metagenomics and metatranscriptomics additionally allow researchers to explore ﻿taxon-function relationships, which can deliver further information about ecosystems. Recent studies address these advantages and their implementation in freshwater assessments (Cordier et al., 2020; Leese et al., 2018).

Metatranscriptomics traditionally involves an mRNA enrichment step to investigate actively expressed genes in the form of coding RNA. However, it is possible to skip the mRNA enrichment step and to investigate the total RNA, including non-coding rRNA. This approach was originally referred to as double-RNA approach (Urich et al., 2008) and later as metatranscriptomics analysis of total rRNA (Turner et al., 2013), total RNA sequencing (RNA-Seq) (Bang-Andreasen et al., 2020; Li et al., 2016; Li and Guan, 2017), ﻿total RNA-based metatranscriptomics, and ﻿total RNA-seq-based metatranscriptomics (Li and Guan, 2017). To distinguish this approach from traditional metatranscriptomics, we will use the term total RNA-Seq in the following. Total RNA-Seq might be advantageous for taxonomic profiling of especially the active portion of microbial communities (Bang-Andreasen et al., 2020; Lanzén et al., 2011; Urich et al., 2008). This assumption is based on two characteristics:

1) Total RNA-Seq naturally enriches sequencing data for barcode sequences of genes that are commonly used for taxonomic profiling. This natural enrichment is achieved because 80-98% of RNA consists of ribosomal RNA (rRNA) (Peano et al., 2013; Westermann et al., 2012) containing the common small subunit (SSU) and large subunit (LSU) rRNA barcodes for prokaryotes (16S and 23S rRNA) and microbial eukaryotes (18S and 28S rRNA). These rRNA barcodes can therefore make up 37% – 71% of total RNA-Seq reads (Elekwachi et al., 2017; Yu and Zhang, 2012). In contrast, metagenomics targets all genes, including genes that are functionally important but contain little taxonomic information, and as a consequence, SSU and LSU rRNA barcodes can make up as little as 0.05% – 1.4% of total reads in metagenomic sequencing data (Logares et al., 2014; Yilmaz et al., 2011). Therefore, the natural enrichment of SSU and LSU rRNA sequences in total RNA-Seq enables a better sequencing coverage of common microbial barcodes in comparison to metagenomics. This might allow for more accurate taxonomic profiling of diverse communities since these sequences are much better represented and taxonomically annotated in public databases than most other parts of the genome. However, this taxonomic advantage comes at the cost of functional coverage, since other genes will be poorly represented and likely better covered by metagenomics.

2) Metatranscriptomics, including total RNA-Seq, can be used to identify the active part of a community through actively expressed RNA (Geisen et al., 2015; Gomez-Silvan et al., 2018). In comparison, metagenomics targets not only the active part of a community but also the DNA of dead and inactive cells and extracellular DNA, which can make up 40-90% of the total DNA pool (Carini et al., 2016; Torti et al., 2015). Consequently, metatranscriptomics might generate more relevant information for ecological assessments as it reflects the portion of the community that is actively interacting with the environment and, therefore, might better reflect environmental conditions. This idea is supported by studies based on metabarcoding, which compared DNA and RNA in terms of their correlation to the environment and in general showed stronger correlations based on RNA (Laroche et al., 2016; Pawlowski et al., 2014; Pochon et al., 2015; Visco et al., 2015). In contrast, metagenomics ﻿could give a broader insight into overall community structures and functions, since it targets all genes of all present organisms.

In this study, we compare metagenomics and total RNA-Seq in terms of their taxonomic profiling performance for microbial communities. Several studies compare the taxonomic composition of environmental microbial communities between total RNA-Seq and metabarcoding (Lanzén et al., 2011; Yan et al., 2018) respectively metagenomics (Lanzén et al., 2011; Shi et al., 2011; Urich et al., 2014). However, to our knowledge, a controlled, mock community-based comparison of total RNA-Seq and metagenomics for taxonomic profiling of microbial communities is lacking. This includes extensive testing of bioinformatics pipelines, which is essential since results from sequencing data are heavily dependent on the bioinformatics tools that are used for their processing (Bashiardes et al., 2016; Knight et al., 2018; McIntyre et al., 2017; Quince et al., 2017; Shakya et al., 2019; Vollmers et al., 2017)\*. Such a comparison could, however, have important implications for ecological assessments of freshwaters and other ecosystems, given the above-mentioned bioindication potential of microbes. Therefore, we applied total RNA-Seq and metagenomics on two sample sets: 1) a commercially available microbial mock community consisting of eight bacterial and two yeast species at log-distributed abundances, and 2) a display tank water sample to simulate environmental freshwater sampling. In addition, we applied and compared a broad variety of common bioinformatic processing and analysis programs to both metagenomics and total RNA-Seq data of both sample sets to test the impact of these programs on taxonomic profiling performance. We generated a pipeline for each combination of bioinformatics programs (1,536 pipelines in total), and we refer to each combination of pipeline and sample type (metagenomics or total RNA-Seq) as a workflow. Ultimately, we tested 1,536 pipelines on each sample type, leading to 3,072 workflows that were applied to each sample set (mock community and display tank).

Our study had two objectives: 1) statistically evaluating the taxonomic profiling performance, as in accuracy and precision, of our 3,072 workflows for the mock community, and 2) comparing the taxonomic profiling performance of workflows between the two sample sets (mock community and display tank) by evaluating the precision conservation between identical workflows applied to both sample sets.

This enabled us to determine if metagenomics or total RNA-Seq performed better for our sample sets in terms of taxonomic profiling. Furthermore, we were able to identify both accurate and biased pipelines. Based on our results, we can recommend a workflow that performs well for taxonomic profiling of microbial communities in different sample sets.

# Methods:

* Give standardized information about the samples according to (Yilmaz et al. 2011) MIxS specifications

The study design is summarized in Fig. 1 and further details are given in the following.

Chart, funnel chart

Description automatically generated

Figure 1: Summary of the study design. Two sample sets were obtained by 1) mixing a commercial microbial mock community with water and 2) taking a water sample from a display tank, which contained fish, crustaceans, molluscs, macrophytes, and an established microbial community. Three replicates per sample set were generated and filtered through 0.2 µm filters. DNA and RNA were extracted in parallel and shotgun sequenced on the same MiSeq run, representing two sample types (metagenomics and total RNA-Seq). The sequencing data were processed using 1,536 pipelines as shown in Fig. 2. Each combination of pipeline and sample type is referred to as a workflow, resulting in 3,072 workflows that were applied to both sample sets. These were statistically evaluated by 1) estimating the accuracy and precision for the mock community sample set and heuristically determining the workflows that were closest to the optimal performance, and 2) estimating the precision conservation between both sample sets for these optimal performing workflows.

*Sampling*

For this study, we used a commercially available microbial mock community (ZymoBIOMICS Microbial Community Standard II (Log Distribution); Zymo Research; Irvine; CA U.S.A.). It consists of eight bacterial (three gram-negative and five gram-positive) and two yeast species. They are mixed by the manufacturer to create log-distributed species abundances based on genomic DNA quantities (Tab. 1). The mock community was preserved in DNA/RNA Shield (Zymo Research; Irvine; CA U.S.A.), which inactivates cells while preserving DNA and RNA. We generated three simulated water sample replicates by adding 130 µl of the microbial mock community to 50 mL ultrapure water respectively.

Table 1: Microbial composition of the mock community (modified from the ZymoBIOMICS Microbial Community Standard II (Log Distribution) manual, for more detailed information see manual).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Defined Composition (%) | | | | |
| Genomic DNA | 16S SSU Only1 | 16S & 18S SSU1 | Genome Copy2 | Cell Number3 |
| *Listeria monocytogenes* | 89.1 | 95.9 | 91.9 | 94.8 | 94.9 |
| *Pseudomonas aeruginosa* | 8.9 | 2.8 | 2.7 | 4.2 | 4.2 |
| *Bacillus subtilis* | 0.89 | 1.2 | 1.1 | 0.7 | 0.7 |
| *Saccharomyces cerevisiae* | 0.89 | NA | 4.1 | 0.23 | 0.12 |
| *Escherichia coli* | 0.089 | 0.069 | 0.066 | 0.058 | 0.058 |
| *Salmonella enterica* | 0.089 | 0.07 | 0.067 | 0.059 | 0.059 |
| *Lactobacillus fermentum* | 0.0089 | 0.012 | 0.012 | 0.015 | 0.015 |
| *Enterococcus faecalis* | 0.00089 | 0.00067 | 0.00064 | 0.001 | 0.001 |
| *Cryptococcus neoformans* | 0.00089 | NA | 0.0014 | 0.00015 | 0.00007 |
| *Staphylococcus aureus* | 0.000089 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |

1 The theoretical composition in terms of 16S (or 16S & 18S) SSU rRNA gene abundance was calculated from theoretical genomic DNA composition with the following formula: 16S/18S copy number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp) × 16S/18S copy number per genome; 2 The theoretical composition in terms of genome copy number was calculated from theoretical genomic DNA composition with the following formula: genome copy number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp); 3 The theoretical composition in terms of cell number was calculated from theoretical genomic DNA composition with the following formula: cell number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp)/ploidy.

We also took three one-liter water samples from a display tank (Hagen Aqualab; University of Guelph; Guelph; ON Canada) containing various fish, mollusks, crustacean, and macrophyte species (Sup. Tab. 1) as well as an established microbial community to simulate environmental freshwater sampling (Sup. Fig. 1) using a bleach-sterilized jug.

*Laboratory processing*

We handled water samples in a separate clean laboratory (for details see Supplemental material 1). All samples were filtered through sterile 0.2 µm Nalgene Analytical Test Filter Funnels (Thermo Fisher Scientific; Burlington; ON Canada) using an 80 mbar Welch WOB-L® Dry Vacuum Pump (VWR International; Mississauga; ON Canada). We filtered both types of samples on two different days. On the first day (31 Jan 2020), we filtered the three 50 mL microbial mock community mixtures and added a negative filtration control by additionally filtering 50 mL of the ultrapure water that was used to set up the mixtures. On the second day (05 Feb 2020), we filtered three times 1 L of the display tank water and added a negative filtration control by adding another dry filter to the subsequent protocol. After each filtration, we immediately cut filters into small pieces and transferred them into ZR BashingBead Lysis Tubes (0.1 & 0.5 mm) (Zymo Research; Irvine; CA U.S.A.) which were prepared with 1 mL of DNA/RNA Shield under a clean hood in a low DNA-concentration laboratory before filtration.

BashingBead tubes were shaken using a Vortex-Genie 2 (Scientific Industries, Inc.; Burlington; NY U.S.A.) with a Horizontal-(24) Microtube holder (Scientific Industries, Inc.; Burlington; NY U.S.A.) for 40 min at maximum rpm to break up cells following manufacturer’s instructions for optimal cell breakup of the purchased mock community.

For parallel DNA/RNA extraction from samples, we used a modified version of the Quick-DNA/RNA Microprep Plus Kit (Zymo Research; Irvine; CA U.S.A.). We added a purification step using Zymo-Spin II‑µHRC Filters (Zymo Research; Irvine; CA U.S.A.) and modified the protocol to process more sample volume (for details see Supplemental material 2). We extracted the two types of samples, including the negative filtration controls, on two different days under a clean hood in a low DNA-concentration laboratory and added a negative extraction control each day by processing only the extraction buffer along with the other samples. That way, we accounted for possible contamination during the filtrations as well as during the extractions.

Extracted DNA and RNA along with all negative filtration and extraction controls were sent to Génome Québec (Montreal; QC Canada) for library preparation and shotgun sequencing. Processing steps and quality control of RNA samples were given by the sequencing centre as follows:

Total RNA was quantified using a NanoDrop Spectrophotometer ND-1000 (NanoDrop Technologies, Inc.) (Sup. Tab. 2) and its integrity was assessed on a 2100 Bioanalyzer (Agilent Technologies) (Sup. Tab. 3). Libraries were generated from 1 µL of each sample as follows: cDNA synthesis was achieved with the NEBNext RNA First Strand Synthesis E7771 and NEBNext Ultra Directional RNA Second Strand Synthesis Modules (New England Biolabs; Whitby; ON Canada). The remaining steps of library preparation were done using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs; Whitby; ON Canada). Adapters and PCR primers were purchased from New England Biolabs (Whitby; ON Canada). Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Roche Sequencing Solutions Inc; Pleasanton; CA U.S.A). Average fragment size was determined using a LabChip GXII instrument (PerkinElmer). Note that the mRNA enrichment step was skipped to create total RNA libraries.

gDNA was quantified using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Thermo Fisher Scientific; Burlington; ON Canada) (Sup. Tab. 4). Libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs; Whitby; ON Canada) as per the manufacturer’s recommendations. Adapters and PCR primers were purchased from IDT (Coralville; IA U.S.A.). Size selection of libraries for the desired insert size was performed using SparQ beads (VWR; Mississauga; ON Canada). Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Roche Sequencing Solutions Inc; Pleasanton; CA U.S.A). Average fragment size was determined using a LabChip GXII instrument (PerkinElmer).

5 µL of both the DNA and RNA libraries were respectively combined and used for quality control. Afterwards, 15 µL of both the DNA and RNA library pool were combined and used for quality control and sequencing.

During library preparation, normalization was performed by processing equal volumes of samples instead of the common procedure to process equal concentrations of samples. We chose this alternative normalization method because it allowed for an equal relative sequencing depth per sample as opposed to an equal total sequencing depth. That way, the relative numbers of reads per sample mirrored the relative amount of DNA/RNA in each sample, avoiding an over- or underrepresentation of samples with higher or lower DNA/RNA amounts. The DNA libraries yielded fragments of ~438 bp length, whereas the RNA libraries yielded fragments of ~303 bp length (both including adaptors and indices). To be able to compare DNA and RNA without choosing unnecessarily long paired-end reads, both libraries were sequenced in a single Illumina MiSeq PE 150 bp run.

*Bioinformatic processing*

﻿ We obtained XXX paired-end reads (Bioproject number: XXX, SRA accession number: XXX), with an average of XXX per sample, whereby the display tank RNA samples contained on average an order of magnitude more sequences than the other samples due to our normalization method (Sup. Tab. 5+6 – note: will be turned into sunburst diagram). We processed the sequences in six steps, using multiple common tools/parameters for each step (Fig. 2).

Step one (trimming and quality filtering):

Applied PHRED score cut-offs for trimming and quality filtering of High-Throughput Sequencing data vary across the literature. While strict quality trimming, i.e., trimming at high PHRED score cut-offs of 20-30, is common (Deiner et al., 2017; MacManes, 2014), gentle quality trimming at PHRED score cut-offs of ﻿2-﻿5 can improve transcriptome assembly (MacManes, 2014). To explore the effect of different PHRED score cut-offs on both metagenomics and total RNA-Seq taxonomic profiling performance, we used Trimmomatic v0.39 (Bolger et al., 2014) at four different PHRED score cut-offs (﻿PHRED ≤5, ≤10, ≤15, and ≤20). We trimmed the leading and trailing low-quality nucleotides of each read and ran a sliding window of size 4 over each read, cutting it if the average quality of nucleotides in the sliding window was below the respective PHRED score cut-offs. Additionally, we excluded reads shorter than 25 nucleotides after trimming. The exact command was the following, where X represents the respective PHRED score cut-off: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10, LEADING:X TRAILING:X, SLIDINGWINDOW:4:X MINLEN:25. After trimming, we error-corrected all reads using the error-correction module of the assembler SPAdes v 3.14.1 (Bankevich et al., 2012) by running SPAdes on forward and reverse reads with the parameter --only-error-correction. That way, error-correction between the assemblers was standardized.

Step two (rRNA sorting):

We used three approaches to sort reads into rRNA and non-rRNA reads: 1) alignment-based with SortMeRNA v4.0.0 (Kopylova et al., 2012), which sorts reads by aligning them to rRNA databases. Trimmed forward and reverse reads were aligned against all program-internal rRNA databases using parameters ‑fastx to generate output files in fasta format, ‑num\_alignments 1 to only filter the reads, --paired\_in to keep both forward and reverse reads if only one matches, --out2 to save forward and reverse reads in separate files, and all other parameters set to default; resulting read names were concatenated and reads were extracted from the trimmed forward and reverse reads to obtain paired reads, 2) Hidden Markov model-based (HMM-based) with barrnap v0.9 (Seemann, unpublished), which predicts the location of rRNA genes in genomes using trained HMMs. However, since barrnap only keeps reads that contain rRNA genes, we used it here to identify rRNA-reads. Both trimmed forward and reverse reads were separately run against all three domains of life (--kingdom euk, arc, and bac) using parameters --lencutoff 0.000001 and --reject 0.000001 to keep partial matches, and all other parameters set to default. The read names of both filtered read files were concatenated and reads were extracted from the trimmed forward and reverse reads to obtain paired reads, and 3) kmer-based with rRNAFilter v1.1 (Wang et al., 2017), which filters reads based on the difference of the k-mer frequencies of rRNA and non-rRNA reads. Since rRNA reads are much more abundant, their k-mer frequencies are distinguishable by rRNAFilter. Both trimmed forward and reverse reads were separately filtered with default parameters including k=20. The read names of both filtered read files were concatenated and reads were extracted from the trimmed forward and reverse reads to obtain paired reads.

For these three approaches, we excluded non-rRNA-reads thereafter. Additionally, we performed a *no sorting* approach where we did not sort reads but used all reads, leading to four approaches in total.

Step three (assembly):

We used eight scaffold assemblers on both DNA and RNA reads, four assemblers optimized for DNA: 1) SPAdes using default parameters, 2) metaSPAdes v3.14.1 (Nurk et al., 2017) using default parameters, 3) MEGAHIT v1.2.9 (Li et al., 2015) using the parameter ‑‑presets meta-large, which changes the used k-mers to better assemble large and complex metagenomes, and otherwise default parameters, 4) IDBA-UD v1.1.1 (Peng et al., 2012) using default parameters; and four assemblers optimized for RNA: 5) Trinity v2.10.0 (Grabherr et al., 2013) using default parameters, 6) rnaSPAdes v3.14.1 (Bushmanova et al., 2019) using default parameters, 7) IDBA-tran v1.1.1 (Peng et al., 2013) using default parameters, and 8) Trans-ABySS v2.0.1 (Robertson et al., 2010) using default parameters. We chose to test both DNA- and RNA-optimized assemblers for DNA and RNA reads since total RNA-Seq read composition is different from traditional metatranscriptomics read composition, and we wanted to test how both assembler types deal with this specific read composition. Note that all assemblers but Trans-ABySS run multiple k-mers by default, whereas Trans-ABySS runs only with one k-mer by default.

Step four (mapping individual sample reads back to scaffolds):

We used two programs to map forward and reverse reads back to assembled scaffolds to determine the read abundance of each scaffold: 1) BWA v0.7.17 (Li and Durbin, 2009) using default parameters and 2) Bowtie2 v2.3.3.1 (Langmead and Salzberg, 2012) using default parameters.

Step five (reference database):

We used two reference databases for taxonomic annotation: 1) NCBI nt (Agarwala et al., 2016), downloaded on 03 Feb 2020, and 2) SILVA132\_NR99 (Quast et al., 2013), downloaded on 28 Aug 2020. For the latter, we downloaded both the available SSU and LSU sets and concatenated them into one SILVA database.

Step six (taxonomic annotation):

Taxonomic annotation of scaffolds can be performed with numerous tools that fit into different categories based on the methods they use, such as sequence similarity methods and sequence composition methods, which require reference databases, phylogenetic methods, and probabilistic methods. We limited our taxonomic annotation to tools based on sequence similarity methods and explored two common tools, kraken2 v2.1.1 (Wood et al., 2019) based on k-mer matching and BLAST v2.10.0 (Altschul et al., 1990) based on local alignments. Since BLAST is highly dependent on parameter choice and processing of the results, we applied it in two different ways, therefore using three approaches in total to taxonomically classify each scaffold: 1) k-mer-based with kraken2 using default parameters, 2) alignment-based with BLAST with an E-value cut-off of e‑05, otherwise default parameters, and keeping the taxonomy hit with the highest bitscore per sequence unless there were multiple taxonomy hits with an identical highest bitscore, in which case the lowest common ancestor (LCA) was identified, and 3) alignment-based with BLAST with an E-value cut-off of e-05, otherwise default parameters and further filtering of the taxonomy hits. Filtering was carried out by mimicking steps of the programs CREST (Lanzén et al., 2012) and BASTA (Kahlke and Ralph, 2019): filtering out taxonomy hits below a bitscore of 155 and an alignment length of 100, only keeping taxonomy hits within 2% of the best bitscore of each sequence, applying a cut-off for taxonomic ranks based on BLAST pident values (species: 99%, genus: 97%, family: 95%, order: 90%, class: 85%, phylum: 80%), and identifying the LCA.

To be able to compare the taxonomic annotation results based on SILVA and NCBI nt, we standardized the taxonomy by translating it for all SILVA hits into NCBI staxids. \*

We generated a script to run a pipeline for all possible combinations of steps and programs (1,536 pipelines). The full code including all program parameters and versions, as well as the scripts for translating SILVA taxonomy to NCBI staxids, for creating SILVA BLAST and kraken2 databases, and for filtering BLAST results based on CREST and BASTA are available on GitHub (XXX). Each pipeline resulted in a table of assembled scaffolds, taxonomic annotations, and absolute read counts, for a total of 1,536 tables per sample.

Diagram

Description automatically generated

Figure 2: Summary of the steps and tools used to process the sequencing data. Each combination of tools is referred to as a pipeline, resulting in 1,536 pipelines. The combination of each pipeline and sample type (metagenomics and total RNA-seq) is referred to as a workflow, resulting in 3,072 workflows that were applied to each sample set (mock community and display tank). Note that in step 3, some assemblers are DNA-optimized and some RNA-optimized, yet we tested all on both metagenomics and total RNA-Seq data. We chose this approach since total RNA-Seq read composition is different from traditional metatranscriptomics read composition, and we wanted to test how both assembler types deal with this specific read composition.

*Statistical evaluation*

For the statistical evaluation, all tables were further processed in R (VERSION). For each pipeline, scaffolds with identical taxonomic annotation were summarized and their absolute read count added up. The combination of each pipeline and sample type (metagenomics or total RNA-Seq) is referred to as a workflow.

Our first objective was to statistically evaluate the performance of our workflows on the mock community samples. Therefore, we calculated both accuracy and precision for each pipeline utilized for the metagenomics and total RNA-Seq samples. Accuracy was calculated using Chi-square tests for each workflow, taking into account the observed and expected abundance of each species in the mock community and additional, false-positive species that could have been introduced through the pipelines. The expected abundance of each species in the mock community was independently determined for each sample by multiplying the total number of reads of each sample with the relative cell abundance of each species given in Tab. 1. Species that were introduced through the pipelines were assigned with an expected abundance of 0. The observed abundance of each species is given by their absolute read count for each workflow. To calculate precision, we determined the variance of each workflow across the three replicates.

We plotted the accuracy of all workflows against precision. We heuristically selected the X workflows that were closest to the optimal performance, i.e., the origin of the plot, and determined them as the best performing workflows for our mock community samples.

To test if the workflows that performed best for the mock communities performed similarly well for the display tank samples, our second objective was to compare the performance of the selected workflows between the two sample sets (mock community and display tank samples). Therefore, we determined the workflow precision for the display tank samples by calculating the variance across the three replicates.

We plotted the precision for the mock community samples against the precision for the display tank samples. The workflows whose points fell closest to the identity line of the plot had the highest levels of precision conservation between the two sample sets. We assumed that if the precision of a workflow was conserved between both sample sets, its accuracy was also conserved. Consequently, we heuristically identified X workflows that had the highest combination of precision and accuracy for both the mock community and display tank samples.

The full R code with all pack versions is available as Jupyter Notebook on GitHub (XXX).

The other methods we’ll use (for example PCA) will be determined on the go so not possible to describe them here yet but will be added in the future

# Results

Do DNA assemblers work better on DNA and RNA assemblers better on RNA?

How close are different pipeline results to “Gold standard community”?

Which process steps cause no significant differences?

# Discussion:

Shotgun sequencing will highly favour microbes due to their high abundance, and metazoans living in an ecosystem could stay undetected.

Point out that results are valid only for this specific sequencing depth. If sequencing depth increases, metagenomics will become more effective (because it doesn’t only cover rRNA barcodes but many more), but we don’t know to what degree/at what depth in relation to community complexity.

RNA hard to handle, short half-life, not feasible for biomonitoring? Reference Cordier. Stabilization 🡪 how effective?

Taxonomy-free biomonitoring

Total RNA Seq can be even further used for mRNA analysis – two birds with one stone

Role of mRNA-Seq for taxonomic profiling? (need to find REFS)

Focus on microbes but also works for other organisms (context eDNA)

Awareness that programs could have been optimized more, for example different k-mers for assemblers.

Awareness that mock community is well represented in NCBI and might work better for metagenomics as compared to display sample

However, the application of metagenomics for taxonomic profiling is highly dependent on sequencing depth, i.e., the sequenced portion of a sample. This dependency is due to the small number of genes in an organism that are useful for taxonomic profiling. Standard barcode genes that are used for that purpose can make up as little as 0.05%-1.4% in metagenomics data (Logares et al., 2014; Yilmaz et al., 2011), and references for other genes are often missing, which is why the major portion of metagenomic data often remains unknown (Singer et al., 2020; Stat et al., 2017). The efficiency of metagenomics for taxonomic profiling is, therefore, dependant on the coverage of these barcode genes, which is directly related to the sequencing depth. The more complex a community is, the higher sequencing depth is required to cover barcode genes to an extend where the community can be accurately profiled. Consequently, given sufficient sequencing depth, metagenomics can outperform metabarcoding in terms of taxonomic profiling (Logares et al., 2014; Shah et al., 2010; Shakya et al., 2013), whereas at low sequencing depths, the coverage of barcodes can sometimes be too low and metagenomics can be outperformed by metabarcoding (Singer et al., 2020; Stat et al., 2017).

Another factor that needs to be considered for metagenomics is the high costs that are involved for high sequencing depths. Given the previously mentioned dependency of metagenomics on sequencing depth, it is recommended to aim for ﻿maximized sequencing output when performing metagenomics. ﻿Illumina HiSeq, NextSeq, and NovaSeq sequencing runs are considered to be suited for that purpose (Quince et al., 2017). However, running these sequencers is expensive, and for many biomonitoring applications, it is important to consider the limited budget available to stakeholders. Consequently, the use of these high sequencing output sequencers might not be affordable yet for routine biomonitoring.

We predicted that by using total RNA-Seq, we would be able to accurately profile the entire microbial mock community at a moderate sequencing depth and that this approach would outperform metagenomics. This would make total RNA-Seq attractive and affordable for stakeholders. We also predicted that significant differences would be observable for both the microbial mock community and the display tank sample using different bioinformatic pipelines.

# Acknowledgements

Sarah Adamowicz, Karl Cottenie, Nicole Ricker, Anders Lanzen

# References

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., et al. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 44, D7–D19. doi:10.1093/nar/gkv1290.

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., and Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* 9, 134–147. doi:10.1111/2041-210X.12849.

Almeida, O. G. G., and De Martinis, E. C. P. (2019). Bioinformatics tools to assess metagenomic data for applied microbiology. *Appl. Microbiol. Biotechnol.* 103, 69–82. doi:10.1007/s00253-018-9464-9.

Altschul, stephen F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403–410.

Bang-Andreasen, T., Anwar, M. Z., Lanzén, A., Kjøller, R., Rønn, R., Ekelund, F., et al. (2020). Total RNA sequencing reveals multilevel microbial community changes and functional responses to wood ash application in agricultural and forest soil. *FEMS Microbiol. Ecol.* 96, 1–13. doi:10.1093/femsec/fiaa016.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021.

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

Bellinger, E. G., and Sigee, D. C. (2015). *Freshwater Algae - Identification, Enumeration and Use as Bioindicators*. 2nd ed. Chichester, West Sussex: John Wiley & Sons Ltd.

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

Burger, J. (2006). Bioindicators: A review of their use in the environmental literature 1970–2005. *Environ. Bioindic.* 1, 136–144. doi:10.1080/15555270600701540.

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

Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S., and Fierer, N. (2016). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* 2, 1–6. doi:10.1038/nmicrobiol.2016.242.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., and Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Sci. Adv.* 1. doi:10.1126/sciadv.1400253.

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., and Pawlowski, J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends Microbiol.* 27, 387–397. doi:10.1016/j.tim.2018.10.012.

Cordier, T., Sáez, L. A., Apotheloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems Monitoring Powered by Environmental Genomics: A Review of Current Strategies with An Implementation Roadmap. *Preprints*. doi:10.20944/PREPRINTS202001.0278.V1.

Corse, E., Tougard, C., Archambaud-Suard, G., Agnèse, J. F., Messu Mandeng, F. D., Bilong Bilong, C. F., et al. (2019). One-locus-several-primers: A strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding studies. *Ecol. Evol.* 9, 4603–4620. doi:10.1002/ece3.5063.

Costanza, R., de Groot, R., Sutton, P., van der Ploeg, S., Anderson, S. J., Kubiszewski, I., et al. (2014). Changes in the global value of ecosystem services. *Glob. Environ. Chang.* 26, 152–158. doi:10.1016/j.gloenvcha.2014.04.002.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., et al. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Mol. Ecol. Resour.* 14, 306–323. doi:10.1111/1755-0998.12188.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi:10.1111/mec.14350.

Dextrase, A. J., and Mandrak, N. E. (2006). Impacts of alien invasive species on freshwater fauna at risk in Canada. *Biol. Invasions* 8, 13–24. doi:10.1007/s10530-005-0232-2.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., et al. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biol. Rev. Camb. Philos. Soc.* 81, 163–182. doi:10.1017/S1464793105006950.

Elbrecht, V., and Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS One* 10, 1–16. doi:10.1371/journal.pone.0130324.

Elekwachi, C. O., Wang, Z., Wu, X., Rabee, A., and 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. *Front. Microbiol.* 8, 1–14. doi:10.3389/fmicb.2017.01814.

Foissner, W., and Berger, H. (1996). A User-Friendly Guide to the Ciliates. *Freshw. Biol.* 35, 375–482.

Geisen, S., Tveit, A. T., Clark, I. M., Richter, A., Svenning, M. M., Bonkowski, M., et al. (2015). Metatranscriptomic census of active protists in soils. *ISME J.* 9, 2178–2190. doi:10.1038/ismej.2015.30.

Gomez-Silvan, C., Leung, M. H. Y., Grue, K. A., Kaur, R., Tong, X., Lee, P. K. H., et al. (2018). A comparison of methods used to unveil the genetic and metabolic pool in the built environment. *Microbiome* 6, 1–16. doi:10.1186/s40168-018-0453-0.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* 29, 644–652. doi:10.1038/nbt.1883.Trinity.

Haury, J., Peltre, M.-C., Trémolières, M., Barbe, J., Thiébaut, G., Bernez, I., et al. (2006). A new method to assess water trophy and organic pollution – the Macrophyte Biological Index for Rivers (IBMR): its application to different types of river and pollution. *Hydrobiologia* 570, 153–158. doi:10.1007/s10750-006-0175-3.

IPBES (2019). *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. , eds. S. Díaz, J. Settele, E. S. Brondízio, H. T. Ngo, M. Guèze, J. Agard, et al. Bonn, Germany Available at: https://doi.org/10.5281/zenodo.3553579.

Jane, S. F., Wilcox, T. M., Mckelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., et al. (2015). Distance, flow and PCR inhibition: EDNA dynamics in two headwater streams. *Mol. Ecol. Resour.* 15, 216–227. doi:10.1111/1755-0998.12285.

Jensen, K., Trepel, M., Merritt, D., and Rosenthal, G. (2006). Restoration ecology of river valleys. *Basic Appl. Ecol.* 7, 383–387. doi:10.1016/j.baae.2006.05.008.

Kahlke, T., and Ralph, P. J. (2019). BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* 10, 100–103. doi:10.1111/2041-210X.13095.

Karr, J. R. (1981). Assessment of Biotic Integrity Using Fish Communities. *Fisheries* 6, 21–27. doi:10.1577/1548-8446(1981)006.

Knight, R., Navas, J., Quinn, R. A., Sanders, J. G., and Zhu, Q. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* doi:10.1038/s41579-018-0029-9.

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

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., and Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci. Rep.* 7, 1–12. doi:10.1038/s41598-017-17333-x.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

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

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., et al. (2012). CREST - Classification Resources for Environmental Sequence Tags. *PLoS One* 7. doi:10.1371/journal.pone.0049334.

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lejzerowicz, F., Pawlowski, J., et al. (2016). First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. *Mar. Environ. Res.* 120, 225–235. doi:10.1016/j.marenvres.2016.08.009.

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., et al. (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. *Adv. Ecol. Res.* 58, 63–99. doi:10.1016/bs.aecr.2018.01.001.

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

Li, F., and Guan, L. L. (2017). Metatranscriptomic profiling reveals linkages between the active rumen microbiome and feed efficiency in beef cattle. *Appl. Environ. Microbiol.* 83, 1–16. doi:10.1128/AEM.00061-17.

Li, F., Henderson, G., Sun, X., Cox, F., Janssen, P. H., and Guan, L. L. (2016). Taxonomic assessment of rumen microbiota using total rna and targeted amplicon sequencing approaches. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.00987.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* 16, 2659–2671. doi:10.1111/1462-2920.12250.

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Front. Genet.* 5, 1–7. doi:10.3389/fgene.2014.00013.

Malmqvist, B., and Rundle, S. (2002). Threats to the running water ecosystems of the world. *Environ. Conserv.* 29, 134–153. doi:10.1017/S0376892902000097.

McArthur, J. V (2001). “Bacteria as Biomonitors,” in *Bioassessment and Management of North American Freshwater Wetlands*, eds. R. B. Rader, D. P. Batzer, and S. A. Wissinger (Chichester, West Sussex: John Wiley & Sons), 249–261. doi:https://doi.org/10.1002/aqc.509.

McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., et al. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* 18, 1–19. doi:10.1186/s13059-017-1299-7.

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., et al. (2018). Minimizing polymerase biases in metabarcoding. *Mol. Ecol. Resour.* 18, 927–939. doi:10.1111/1755-0998.12895.

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

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., et al. (2012). CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biol.* 10, e1001419. doi:10.1371/journal.pbio.1001419.

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

Pawlowski, J., Lejzerowicz, F., and Esling, P. (2014). Next-generation environmental diversity surveys of foraminifera: Preparing the future. *Biol. Bull.* 227, 93–106. doi:10.1086/BBLv227n2p93.

Payne, R. J. (2013). Seven reasons why protists make useful bioindicators. *Acta Protozool.* 52, 105–113. doi:10.4467/16890027AP.13.0011.1108.

Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., et al. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Inform. Exp.* 3, 1–11. doi:10.1186/2042-5783-3-1.

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

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

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., et al. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science (80-. ).* 344. doi:10.1126/science.1246752.

Piñol, J., San Andrés, V., Clare, E. L., Mir, G., and Symondson, W. O. C. (2014). A pragmatic approach to the analysis of diets of generalist predators: The use of next-generation sequencing with no blocking probes. *Mol. Ecol. Resour.* 14, 18–26. doi:10.1111/1755-0998.12156.

Piñol, J., Senar, M. A., and Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Mol. Ecol.* 28, 407–419. doi:10.1111/mec.14776.

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., et al. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *Gigascience* 8, 1–22. doi:10.1093/gigascience/giz092.

Pochon, X., Wood, S. A., Keeley, N. B., Lejzerowicz, F., Esling, P., Drew, J., et al. (2015). Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Mar. Pollut. Bull.* 100, 370–382. doi:10.1016/j.marpolbul.2015.08.022.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* 41, 590–596. doi:10.1093/nar/gks1219.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi:10.1038/nbt.3935.

Resh, V. H., and Unzicker, J. D. (1975). Water Quality Monitoring and Aquatic Organisms : The Importance of Species Identification. *Water Pollut. Control Fed.* 47, 9–19.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912. doi:10.1038/nmeth.1517.

Seemann, T. BAsic Rapid Ribosomal RNA Predictor - barrnap.

Shah, N., Tang, H., Doak, T. G., and Ye, Y. (2010). Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pacific Symp. Biocomput. 2011*, 165–176.

Shakya, M., Lo, C. C., and Chain, P. S. G. (2019). Advances and challenges in metatranscriptomic analysis. *Front. Genet.* 10, 1–10. doi:10.3389/fgene.2019.00904.

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* 15, 1882–1899. doi:10.1111/1462-2920.12086.

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

Singer, G. A. C., Shekarriz, S., McCarthy, A., Fahner, N., and Hajibabaei, M. (2020). The utility of a metagenomics approach for marine biomonitoring. *bioRxiv*, 2020.03.16.993667. doi:10.1101/2020.03.16.993667.

Smith, M. B., Rocha, A. M., Smillie, C. S., Olesen, S. W., Paradis, C., Wu, L., et al. (2015). Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *MBio* 6, e00326-15. doi:10.1128/mBio.00326-15.

Stat, M., Huggett, M. J., Bernasconi, R., Dibattista, J. D., Berry, T. E., Newman, S. J., et al. (2017). Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. *Sci. Rep.* 7, 1–11. doi:10.1038/s41598-017-12501-5.

Stein, E. D., White, B. P., Mazor, R. D., Jackson, J. K., Battle, J. M., Miller, P. E., et al. (2014). Does DNA barcoding improve performance of traditional stream bioassessment metrics? *Freshw. Sci.* 33, 302–311. doi:10.1086/674782.

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., and Pawlowski, J. (2018). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecol. Indic.* 85, 153–164. doi:10.1016/j.ecolind.2017.10.041.

Sweeney, B. W., Battle, J. M., Jackson, J. K., and Dapkey, T. (2011). Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *J. North Am. Benthol. Soc.* 30, 195–216. doi:10.1899/10-016.1.

Sze, M. A., and Schloss, P. D. (2019). The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. *mSphere* 4, e00163-19. doi:10.1128/msphere.00163-19.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2012.05470.x/full%5Cnpapers2://publication/uuid/30F6E470-48F9-4C24-A3C1-6964EE26B34F.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., et al. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res.* 24, 3189–3194.

Torti, A., Lever, M. A., and Jørgensen, B. B. (2015). Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar. Genomics* 24, 185–196. doi:10.1016/j.margen.2015.08.007.

Turner, T. R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., et al. (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME J.* 7, 2248–2258. doi:10.1038/ismej.2013.119.

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

Urich, T., Lanzén, A., Stokke, R., Pedersen, R. B., Bayer, C., Thorseth, I. H., et al. (2014). Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environ. Microbiol.* 16, 2699–2710. doi:10.1111/1462-2920.12283.

Visco, J. A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., and Pawlowski, J. (2015). Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. *Environ. Sci. Technol.* 49, 7597–7605. doi:10.1021/es506158m.

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

Wang, Y., Hu, H., and Li, X. (2017). rRNAFilter: A Fast Approach for Ribosomal RNA Read Removal Without a Reference Database. *J. Comput. Biol.* 24, 368–375. doi:10.1089/cmb.2016.0113.

Westermann, A. J., Gorski, S. A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10, 618–630. doi:10.1038/nrmicro2852.

Will, K. W., and Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20, 47–55. doi:10.1111/j.1096-0031.2003.00008.x.

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

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput. Biol.* 6. doi:10.1371/journal.pcbi.1000667.

WWF (2020). *Living Planet Report 2020 - Bending the curve of biodiversity loss*. , eds. R. E. A. Almond, M. Grooten, and T. Petersen Gland, Switzerland.

Yan, Y. W., Jiang, Q. Y., Wang, J. G., Zhu, T., Zou, B., Qiu, Q. F., et al. (2018). Microbial communities and diversities in mudflat sediments analyzed using a modified metatranscriptomic method. *Front. Microbiol.* 9, 1–15. doi:10.3389/fmicb.2018.00093.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420. doi:10.1038/nbt.1823.

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

Zhang, G. K., Chain, F. J. J., Abbott, C. L., and Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evol. Appl.* 11, 1901–1914. doi:10.1111/eva.12694.

# Supplemental material

A picture containing algae, broccoli, food, sitting

Description automatically generated

Supplementary Figure 1: Display tank

**Supplemental material 1: filtration**

**Needed:**

* Zymo Bashing Bead tubes (ZR BashingBead Lysis Tubes (0.1 & 0.5 mm))
* DNA/RNA Shield reagent
* 0.2 µm Nalgene Analytical Test Filter Funnels
* Forceps, scissors, burner
* Bleach + EtOH

**Preparation:**

* Label bashing tubes (1 per sample + 1 filtration control) and load them with 1 mL of DNA/RNA Shield

**Procedure:**

1. Sterilize forceps and scissors with burner, then bleach, then EtOH before each filtration
2. Filter water sample, make sure to keep the filtration unit closed when opening the wrapping, to hold nothing but the sample over the unit, and to stand as far as possible away from the unit while it is open
3. Cut filter into small pieces (forceps and scissors) while filter lays on the filtration unit, make sure to stand as far as possible away from the unit
4. Transfer pieces into bashing tubes until all pieces are covered with DNA/RNA Shield
5. Clean filtration unit adapter with bleach after each filtration
6. Bashing tubes can be stored at -20°C

**Supplemental material 2: parallel DNA+RNA extraction from freshwater samples using the Quick-DNA/RNA Microprep Plus Kit and Zymo-Spin II-µHRC Filters (Zymo Research)**

*Note: maximum of 11 samples + extraction control possible at a time*

**Needed:**

* Access to a -80°C freezer
* Microcentrifuge
* Bead beater à  recommendation from Zymo Research is a Vortex Genie unit with a 24 Microtube holder
* ZymoBIOMICS Quick-DNA/RNA Microprep Plus Kit
* Zymo-Spin II-µHRC Filters
* 100% EtOH (1.5 mL per sample + extraction control)
* Eliminase (or other RNase-eliminating reagent)
* 11 microtube racks (1x Qubit tubes; 1x 2 mL tubes + samples + extra tubes; 2x 1.5 mL tubes; 3x columns; 4x collection tubes) + 1 rack for falcon tubes + 1 ice rack
* Timer

**Preparation (all one day in advance):**

*Note: 1 set refers to 1x the number of samples + extraction control*

* Clean whole extraction hood with bleach, EtOH, Eliminase (this order)
* Clean microtube and falcon racks in the same way and put in hood
* Clean pipettes, tip boxes (1x small, 1x medium, 3x large), and hood waste the same way
* Label tubes and columns in hood and close them after labelling (no detailed labels necessary unless specified)
  + 1 set 2 mL tubes
  + 1 set Zymo-Spin ICXM columns + collection tubes (labelled “DNA”)
  + 1 set Zymo-Spin IC columns + collection tubes (labelled “RNA”)
  + 9 more sets collection tubes (4 sets labelled “DNA”, 6 sets labelled “RNA”)
  + 2 sets 1.5 mL tubes (1 set labelled “DNA”, 1 set labelled “RNA”)
  + 2 sets Zymo-Spin III u-HCR Filter columns + collection tubes (1 set labelled “DNA”, 1 set labelled “RNA”)
  + 2 sets 1.5 mL tubes with detailed labels (1 set labelled “DNA”, 1 set labelled “RNA”)
  + 2 sets Qubit tubes (1 set labelled “DNA”, 1 set labelled “RNA”) + 4 additional Qubit tubes for standards (2 RNA + 2 DNA)
  + 2 additional 1.5 mL tubes for DNase-mix preparation + 2 additional 5 mL tubes for Qubit solution preparation (RNA + DNA)
* Put 100% EtOH and 5 mL tubes in falcon rack
* UV-sterilize everything overnight

**Procedure:**

1. Place filters/bead tubes (thawed) in Microtube holder on Vortex Genie, vortex for 40 min at max speed (never more than 18 tubes at a time, because >18 tubes will slow vortexing and create inaccurate results)
2. Centrifuge bead tubes for 1 min at 13,000 xg and room temperature
3. Transfer as much of the supernatant as possible into 2 mL tubes (ideally don’t transfer any beads)
4. Add 1 volume of **DNA/RNA Lysis Buffer** to samples and vortex

*Note: because we transferred all the supernatant from bead tubes (instead of 400 µl as specified in protocol), the samples are too large to carry out next steps in one go. The next steps have to be repeated, so keep all columns/tubes/mixes etc. First and second rounds are indicated in brackets.*

1. Transfer 800 µl (first round)/rest (second round) of samples into Zymo-Spin ICXM in a collection tube and centrifuge for 30 sec at 13,000 xg and room temperature. **SAFE THE FLOW-THROUGH**
2. Transfer Filter columns into a new collection tube
3. Add 1 volume of **100% EtOH** (800 µl (first round)/600 µl (second round)) to flow-through and mix well (pipette 10x up and down)
4. Transfer sample into Zymo-Spin IC columns in a collection tube and centrifuge for 30 sec at 13,000 xg and room temperature. (Note: Only 800 µl can be transferred at a time, so this step has to be repeated; repetition can be done right away by transferring the rest of the sample into the same columns in a new collection tube and repeating the centrifugation)
5. Transfer filter columns into a new collection tube

*Repeat steps 5-9 using the same columns.*

1. Remove all collection tubes and respective racks that are not needed anymore from hood
2. DNase treatment:
3. Prepare **DNase Reaction Mix**: for number of samples + extraction control + 1 buffer, mix 35 µl **DNA Digestion Buffer** with 5 µl **DNase I**
4. Wash all columns with 400 µl DNA/RNA Wash Buffer and centrifuge for 30 sec at 13,000 xg and room temperature
5. Transfer all columns in new collection tubes
6. Add 40 µl **DNase I Reaction Mix** directly to the column matrix
7. Incubate columns at room temperature for 15 minutes
8. Add 400 µl **DNA/RNA Prep Buffer** to columns and centrifuge for 30 sec at 13,000 xg and room temperature
9. Transfer all columns in new collection tubes
10. Add 700 µl **DNA/RNA Wash Buffer** to columns and centrifuge for 30 sec at 13,000 xg and room temperature
11. Transfer all columns in new collection tubes
12. Add 400 µl **DNA/RNA Wash Buffer** to columns and centrifuge for 2 min at 13,000 xg and room temperature to ensure compete removal of wash buffer
13. Transfer one column at a time into 1.5 mL tube (no detailed labels) and add 22.5 µl **ZymoBIOMICS DNase/RNase-Free Water** directly to column matrix, let stand for 5 minutes, and centrifuge for 30 sec at 13,000 xg and room temperature to elute DNA/RNA in water. (Note: only 8 open 1.5 mL tubes fit into 24-Microcentrifuge at a time)
14. Place Zymo-Spin II u-HRC Filter columns in new collection tubes and add 600 µl **ZymoBIOMICS HRC Prep Solution**. Centrifuge for 3 min at 8,000 xg and room temperature à “prepared filter columns”
15. Transfer one prepared filter column at a time into 1.5 mL tube (detailed labels), transfer eluted DNA/RNA into prepared filter column, and centrifuge for 3 min at 16,000 xg and room temperature. (Note: only 8 open 1.5 mL tubes fit into 24-Microcentrifuge at a time)
16. Aliquot 2.5 µl of each final sample into Qubit tubes for concentration measurement
17. Clean an ice rack with bleach, EtOH, Eliminase, and put final DNA/RNA samples on ice
18. Carry out Qubit measurement with 2 µl of aliquoted samples
19. Store final samples at -80°C

Supplementary Table 1: Animal and macrophyte species present in the display tank when taking samples. Species whose presence was uncertain are marked with a “?”

|  |  |
| --- | --- |
| **Group** | **Species** |
| Crustaceans | *Caridina multidentata,* Ostracoda sp. (?) |
| Fish | *Aplocheilus lineatus* (?), *Cynodonichthys hildebrandi*, *Danio rerio*, *Poecillia salvatoris* |
| Macrophytes | Anubias barteri, Bolbitis heudelotii, Cryptocoryne sp., Echinodoras sp., Limnophila sessiliflora, Rotala rotundifolia, Sagittaria subulata, Vesicularia sp. |
| Mollusks | Physella acuta, Planorbarius corneus, Melanoides tuberculate, Neritidae sp. (?) |

Supplementary Table 2: Nanodrop Quantification (RNA)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **Concentration [ng/µl]** | **Total RNA [ng]** | **260/230** | **260/280** |
| M4\_RNA | 1.87 | 33.66 | 0.36 | 8.19 |
| M5\_RNA | 3.32 | 59.76 | 0.39 | 2.02 |
| M6\_RNA | 2.96 | 53.28 | 0.33 | 3.79 |
| M\_Neg\_RNA | 0.73 | 13.14 | 0.09 | -1.43 |
| M\_Ext\_RNA | 0.43 | 7.74 | 0.08 | 4.22 |
| F4\_RNA | 35.14 | 632.52 | 0.12 | 1.79 |
| F5\_RNA | 33.22 | 597.96 | 0.42 | 2.12 |
| F6\_RNA | 26.52 | 477.36 | 0.12 | 1.81 |
| F\_Neg\_RNA | 8.89 | 160.02 | 0.03 | 0.96 |
| F\_Ext\_RNA | 10.67 | 192.06 | 0.06 | 0.67 |

Supplementary Table 3: Bioanalysis (RNA)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **28S/18S** | **RIN** | **Concentration [ng/µl]** | **Total RNA [ng]** |
| M4\_RNA | 1.279624 | N/A | 1.87 | 33.66 |
| M5\_RNA | 1.1748 | N/A | 3.32 | 59.76 |
| M6\_RNA | 1.123146 | N/A | 2.96 | 53.28 |
| M\_Neg\_RNA | 0 | 1.2 | 0.73 | 13.14 |
| M\_Ext\_RNA | 0 | 1.7 | 0.43 | 7.74 |
| F4\_RNA | 1.439941 | 7.6 | 35.14 | 632.52 |
| F5\_RNA | 1.366829 | 7.7 | 33.22 | 597.96 |
| F6\_RNA | 0 | N/A | 26.52 | 477.36 |
| F\_Neg\_RNA | 0 | 1.7 | 8.89 | 160.02 |
| F\_Ext\_RNA | 0 | 1.5 | 10.67 | 192.06 |

Supplementary Table 4: Fluorescence Assay Quantification (DNA)

|  |  |  |
| --- | --- | --- |
| **Sample** | **Concentration [ng/µl]** | **Total DNA [ng]** |
| M4\_DNA | 2.7856 | 136.494 |
| M5\_DNA | 3.0507 | 149.484 |
| M6\_DNA | 2.3648 | 115.875 |
| M\_Neg\_DNA | 0 | 0 |
| M\_Ext\_DNA | 0 | 0 |
| F4\_DNA | 32.711 | 1602.839 |
| F5\_DNA | 28.057 | 1374.793 |
| F6\_DNA | 43.549 | 2133.901 |
| F\_Neg\_DNA | 0 | 0 |
| F\_Ext\_DNA | 0.046519 | 2.279 |

Turn this into Sunburst Diagram/Multi-level Pie Chart:

(Note: 4x means 4 rows will be added here: rRNAFilter, SortMeRNA, barrnap, No filter)

Supplementary Table 5: Number of reads and contigs after each pipeline step (DNA)

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Raw reads | Quality filtered reads | rRNA filtered reads |
| M1\_DNA |  | Phred ≤ 5: | *4x…* |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| M2\_DNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| M3\_DNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| F1\_DNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| F2\_DNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| F3\_DNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |

Supplementary Table 6: Number of reads and contigs after each pipeline step (RNA)

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Raw reads | Quality filtered reads | rRNA filtered reads |
| M1\_RNA |  | Phred ≤ 5: | *4x…* |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| M2\_RNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| M3\_RNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| F1\_RNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| F2\_RNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |
| F3\_RNA |  | Phred ≤ 5: |  |
| Phred ≤ 10: |  |
| Phred ≤ 15: |  |
| Phred ≤ 20: |  |