Comparing metagenomics[,] [and ]total RNA metatranscriptomics [and bioinformatics pipelines] for taxonomic profiling accuracy and precision of a microbial mock community and a display tank community [using 1,536 bioinformatics pipelines]

# Abstract

To be added

Keywords:

# Introduction

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, these ecosystems are heavily impacted by agricultural and industrial pollution, habitat fragmentation (wetland drainage, river straightening, and dam building), and the introduction of invasive species (Dextrase and Mandrak, 2006; Jensen et al., 2006). As a result, freshwater ecosystems 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 through 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 diatoms (Bellinger and Sigee, 2015; Haury et al., 2006; Karr, 1981; Resh and Unzicker, 1975); however, there are ongoing efforts to include other microbes (all 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 to 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). Entire communities are 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).

In this study, we compare metagenomics and total RNA metatranscriptomics (total RNA-Seq; metatranscriptomics without an mRNA enrichment step) for taxonomic profiling. The motivation for such a test stems from two potential advantages of metatranscriptomics over metagenomics. \*

First, total RNA-Seq naturally enriches sequencing data for common barcode sequences. 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 barcodes for prokaryotes (16S) and microbial eukaryotes (18S and 28S rRNA), whereas these barcodes can make up as little as 0.05% – 1.4% of total reads in metagenomics sequencing data (Logares et al., 2014; Yilmaz et al., 2011). 16S, 18S rRNA, and 28S rRNA sequences are much better represented and taxonomically annotated in public databases than most other parts of the genome. In theory the natural enrichment of rRNA in total RNA-Seq enables a better sequencing coverage of common microbial barcodes in comparison with metagenomics, which might allow for more accurate taxonomic profiling of taxonomically diverse communities.

Second, metatranscriptomics can be used to identify the active part of a community (Geisen et al., 2015; Gomez-Silvan et al., 2018). In comparison, metagenomics targets not only the active part of a community but also DNA of dead and/or 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 (REF).

To compare the performances of metagenomics and total RNA-Seq in taxonomic profiling accuracy, we applied these techniques 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 accuracy. 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: a) statistically evaluating the taxonomic profiling performance of our 3,072 workflows for the mock community in terms of accuracy and precision, and b) comparing the performance of workflows between the two sample sets (mock community and display tank) by evaluating the precision conservation between workflows applied to both sample sets.

This enabled us to determine if metagenomics or total RNA-Seq performed better 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.

A picture containing text, toiletry, skin cream

Description automatically generated

Figure : 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 was 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 (thee 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. The purchased mock community contained ~1.5 x 109 cells/mL, which means that each of our mock community mixtures contained ~1.95 x 108 cells (Tab. 1).

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Defined Composition (%) | | | | |
| Genomic DNA | 16S Only1 | 16S & 18S1 | 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) 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, mollusk, crustacean, and macrophyte species as well as an established microbial community to simulate environmental freshwater sampling (Supplementary 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/by treating an additional filter the same way as the other filters without actually filtering water. 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 prior to 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 an additional 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. 1) and its integrity was assessed on a 2100 Bioanalyzer (Agilent Technologies) (Sup. Tab. 2). 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. 3). 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. 4+5 – 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):

We used Trimmomatic (Bolger et al., 2014) at four different PHRED scores (﻿PHRED ≤5, ≤10, ≤15, and ≤20) to trim the leading and trailing low-quality nucleotides of each read and to run 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 scores. 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.

Step two (rRNA sorting):

We used three approaches to sort reads into rRNA and non-rRNA reads: 1) alignment-based – SortMeRNA (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 1, ‑num\_alignments 1, --paired\_in, 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) – barrnap (Seemann, unpublished), which predicts the location of rRNA genes in genomes using trained HMMs. However, since barrnap excludes reads that do not contain rRNA genes, we used it here to identify rRNA-reads. Trimmed forward and reverse reads were run against all three domains of life (--kingdom euk, arc, and bac) using parameters --lencutoff 0.000001, --reject 0.000001, 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, and 3) kmer-based – rRNAFilter (Wang et al., 2017), which filters reads based on the difference of the k-mer frequencies of rRNA and non-rRNA reads using k=20 by default. Since rRNA reads are much more abundant, their k-mer frequencies are distinguishable by rRNAFilter. Trimmed forward and reverse reads were filtered with default parameters; resulting read names 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 assemblers, four DNA assemblers, 1) SPAdes (Bankevich et al., 2012) using default parameters, 2) metaSPAdes (Nurk et al., 2017) using default parameters, 3) MEGAHIT (Li et al., 2015) using the parameter ‑‑presets meta-large and otherwise default parameters, 4) IDBA-UD (Peng et al., 2012) using the parameter --pre\_correction and otherwise default parameters, and four RNA assemblers, 5) Trinity (Grabherr et al., 2013) using default parameters, 6) rnaSPAdes (Bushmanova et al., 2019) using default parameters, 7) IDBA-tran (Peng et al., 2013) using the parameter ‑‑pre\_correction and otherwise default parameters, and 8) Trans-ABySS (Robertson et al., 2010) using default parameters.

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

We used two programs, 1) BWA (Li and Durbin, 2009) using default parameters and 2) Bowtie2 (Langmead and Salzberg, 2012) using default parameters.

Step five (reference database):

We used the reference databases NCBI nt (Agarwala et al., 2016), downloaded on 03 Feb 2020, and 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 classification):

We used three approaches: 1) kraken2 (Wood et al., 2019) using default parameters, 2) BLAST (Altschul et al., 1990) with an E-value cut-off of e‑05, otherwise default parameters, and keeping the hit with the highest bitscore per sequence, and 3) BLAST with an E-value cut-off of e-05, otherwise default parameters, and filtering the hits with an in-house script (based on steps performed by the programs CREST (Lanzén et al., 2012) and BASTA (Kahlke and Ralph, 2019): filtering out hits below a bitscore of 155 and an alignment length of 100, only keeping hits within 2% of the best bitscore of each sequence, applying a cut-off for taxonomic ranks based on BLAST pident values, and identifying the lowest common ancestor of each sequence based on the remaining hits).

To be able to compare the taxonomic classification results based on SILVA and NCBI nt, we standardized the taxonomy by translating it for all SILVA hits into NCBI staxids using an in-house script.

We generated a script to run all possible combinations of steps and programs (1536 pipelines). The full code including all program parameters and versions, as well as the scripts for translating SILVA taxonomy to NCBI staxids and for creating SILVA BLAST and kraken2 databases, are available on GitHub (XXX). Each pipeline resulted in a table of assembled scaffolds, taxonomic annotations, and absolute read counts, for a total of 1536 tables per sample.

A picture containing diagram

Description automatically generated

Figure : Bioinformatic processing summary

*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 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 made the assumption 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)

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.

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.

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.

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.

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.

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.

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.

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.

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., 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.

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, 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.

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.

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., 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.

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.

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.

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.

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.

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.

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.

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 : 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 : 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 : 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 : 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 : 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 : 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: |  |