frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 24 November 2023
DOI 10.3389/fmicb.2023.1217750

OPEN ACCESS

EDITED BY
Cristina García-Aljaro,
University of Barcelona, Spain

REVIEWED BY
Drishti Kaul,
J. Craig Venter Institute (La Jolla),
United States
Craig Lee Moyer,
Western Washington University, United States
Anders Lanzén,
Technology Center Expert in Marine and Food Innovation (AZTI), Spain

*CORRESPONDENCE
Christopher A. Hempel
✉ chempel.work@gmail.com
Dirk Steinke
✉ dsteinke@uoguelph.ca

RECEIVED 05 May 2023
ACCEPTED 10 November 2023
PUBLISHED 24 November 2023

CITATION
Hempel CA, Buchner D, Mack L, Brasseur MV, Tulpan D, Leese F and Steinke D (2023) Predicting environmental stressor levels with machine learning: a comparison between amplicon sequencing, metagenomics, and total RNA sequencing based on taxonomically assigned data.
Front. Microbiol. 14:1217750.
doi: 10.3389/fmicb.2023.1217750

Predicting environmental stressor levels with machine learning: a comparison between amplicon sequencing, metagenomics, and total RNA sequencing based on taxonomically assigned data

Christopher A. Hempel[1,2]*, Dominik Buchner[3], Leoni Mack[4], Marie V. Brasseur[5], Dan Tulpan[6,7], Florian Leese[3,8] and Dirk Steinke[1,2]*

[1]Department of Integrative Biology, University of Guelph, Guelph, ON, Canada, [2]Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada, [3]Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany, [4]Faculty of Aquatic Ecology, University of Duisburg-Essen, Essen, Germany, [5]Leibniz Institute for the Analysis of Biodiversity Change, Zoological Research Museum A. Koenig, Bonn, Germany, [6]School of Computer Science, University of Guelph, Guelph, ON, Canada, [7]Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada, [8]Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen, Germany

Introduction: Microbes are increasingly (re)considered for environmental assessments because they are powerful indicators for the health of ecosystems. The complexity of microbial communities necessitates powerful novel tools to derive conclusions for environmental decision-makers, and machine learning is a promising option in that context. While amplicon sequencing is typically applied to assess microbial communities, metagenomics and total RNA sequencing (herein summarized as omics-based methods) can provide a more holistic picture of microbial biodiversity at sufficient sequencing depths. Despite this advantage, amplicon sequencing and omics-based methods have not yet been compared for taxonomy-based environmental assessments with machine learning.

Methods: In this study, we applied 16S and ITS-2 sequencing, metagenomics, and total RNA sequencing to samples from a stream mesocosm experiment that investigated the impacts of two aquatic stressors, insecticide and increased fine sediment deposition, on stream biodiversity. We processed the data using similarity clustering and denoising (only applicable to amplicon sequencing) as well as multiple taxonomic levels, data types, feature selection, and machine learning algorithms and evaluated the stressor prediction performance of each generated model for a total of 1,536 evaluated combinations of taxonomic datasets and data-processing methods.

Results: Sequencing and data-processing methods had a substantial impact on stressor prediction. While omics-based methods detected a higher diversity of taxa than amplicon sequencing, 16S sequencing outperformed all other sequencing methods in terms of stressor prediction based on the Matthews Correlation Coefficient. However, even the highest observed performance for 16S sequencing was still only moderate. Omics-based methods performed poorly overall, but this was likely due to insufficient sequencing depth. Data types had no impact on performance while feature selection significantly improved performance for omics-based methods but not for amplicon sequencing.

Discussion: We conclude that amplicon sequencing might be a better candidate for machine-learning-based environmental stressor prediction than omics-based methods, but the latter require further research at higher sequencing depths to confirm this conclusion. More sampling could improve stressor prediction performance, and while this was not possible in the context of our study, thousands of sampling sites are monitored for routine environmental assessments, providing an ideal framework to further refine the approach for possible implementation in environmental diagnostics.

KEYWORDS

metabarcoding, metatranscriptomics, freshwater, stressor prediction, bioinformatics, ExStream, mesocosm, environmental assessment

1 Background

Globally, ecosystems are experiencing an unprecedented amount of human-induced environmental stress, caused by climate change, land use, pollution, habitat fragmentation, and the introduction of invasive species. As a consequence, ecosystems are deteriorating and biodiversity is declining faster than ever before in human history (Díaz et al., 2019; WWF, 2020; Pettorelli et al., 2021). The loss of biodiversity has extremely negative effects on ecosystem functions and, thereby, ecosystem services, which also reduces the economic value of ecosystems (Kubiszewski et al., 2017). As a consequence, environmental management to protect and restore ecosystems has garnered increased attention, also at the political level (Díaz et al., 2019).

Environmental management includes the identification of prevalent stressors and their impacts on ecosystem health. Microbes (prokaryotes and unicellular eukaryotes) are very good indicators of ecosystem health because they play a crucial role in ecosystems and are extremely sensitive to changes in environmental conditions. Consequently, their community composition can reveal important information about the health and stress levels of ecosystems, which can be utilized for routine biomonitoring to guide measures for the protection and restoration of ecosystems (Smith et al., 2015; Pawlowski et al., 2016; Cordier et al., 2019; Sagova-Mareckova et al., 2021). Microbial community composition is usually determined by using amplicon sequencing, which involves target PCR to amplify taxonomic barcode genes (amplicons), typically the 16S ribosomal RNA (rRNA) gene for prokaryotes, the internal transcribed spacer 2 (ITS-2) 2 for fungi, and the 18S rRNA gene for other microbial eukaryotes. Although this approach can introduce taxonomic and abundance bias due to varying binding affinities and amplification efficiencies of target primers (Pinto and Raskin, 2012; Lozupone et al., 2013; Walker et al., 2015; Meisel et al., 2016; Laursen et al., 2017; Stat et al., 2017), it is widely used because it is comparably cheap and can generate valuable and consistent information on community composition.

In contrast, metagenomics and metatranscriptomics are target-PCR-free methods that are usually applied to analyze the presence and expression of functional genes within communities (Wooley et al., 2010; Bashiardes et al., 2016; Almeida and De Martinis, 2019; Shakya et al., 2019); however, both methods also generate valuable data that can be used for taxonomic identification of community members as an alternative to amplicon sequencing.

Metagenomics targets all DNA in a sample, including non-functional genes, repetitive regions, and genes containing little taxonomic information due to insufficient variation. A vast number of these genes is lacking reference sequences in databases, and therefore, metagenomics generates large amounts of sequences that cannot be taxonomically annotated. At insufficient sequencing depth, this leads to a low biodiversity coverage that is outperformed by that of amplicon sequencing (Yilmaz et al., 2011; Stat et al., 2017; Tessler et al., 2017). However, this limitation can be overcome by increasing the sequencing depth, and if the depth is increased sufficiently, biodiversity coverage through metagenomics can outperform that of amplicon sequencing (Shah et al., 2010; Shakya et al., 2013; Logares et al., 2014; Brumfield et al., 2020).

Total RNA sequencing (total RNA-Seq; Li et al., 2016; Li and Guan, 2017; Bang-Andreasen et al., 2020), also termed double-RNA approach (Urich et al., 2008), metatranscriptomics analysis of total rRNA (Turner et al., 2013), total RNA metatranscriptomics (Xue et al., 2020), or total RNA-seq-based metatranscriptomics (Li and Guan, 2017), refers to metatranscriptomics without an mRNA enrichment step. Cellular RNA consists mostly of rRNA, including 16S and 18S rRNA, which means that a large portion of total RNA-Seq data can be used for taxonomic annotations of microbes. In a previous study, we showed that total RNA-Seq can identify a microbial mock community consisting of 10 species more accurately than metagenomics at almost one order of magnitude lower sequencing depth (Hempel et al., 2022). Therefore, total RNA-Seq combines the advantages of both amplicon sequencing and metagenomics, as it avoids targeted PCR while producing large amounts of 16S and 18S sequences that can be taxonomically annotated.

Both Metagenomics and metatranscriptomics are more costly than amplicon sequencing but they can deliver target-PCR-free functional and taxonomical information across the tree of life, and as a result, there is a growing interest in their application for ecological assessments (Uyaguari-Diaz et al., 2016; Leese et al., 2018; Cordier et al., 2019, 2021).

Another field of research increasingly considered for use in ecological assessments is machine learning. Machine learning comprises algorithms to discover structural patterns in data that can be used to make predictions. Learning, in that sense, means that

the applied algorithms change their behavior through repeated training so that they perform better going forward (Witten and Frank, 2005). Machine learning is increasingly being used in biological sciences, including microbial ecology and environmental assessments, due to its capacity to deal with the expanding scale and complexity of biological data (Ghannam and Techtmann, 2021; Greener et al., 2022). Cordier et al. (2019) stated that machine learning is the most promising approach for routine biomonitoring as it has the potential to be faster, more cost-efficient, and more accurate than current morphology-based methods, and some researchers believe that ecology represents one of the most relevant areas for machine learning because it could solve a wide and diverse variety of ecological problems (Crisci et al., 2012). It already has been applied successfully to amplicon-sequencing-based environmental assessments in freshwater (Smith et al., 2015; Good et al., 2018), marine and coastal water (Cordier et al., 2017, 2018; Gerhard and Gunsch, 2019; Glasl et al., 2019; Frühe et al., 2020; Dully et al., 2021), estuarine sediments (Lanzén et al., 2020), and soil (Hermans et al., 2020), overcoming both the complex biological challenges associated with environmental data and the statistical challenges associated with the interpretation of large datasets. However, for the prediction of ecological variables with taxonomically assigned metagenomic data, machine learning has been applied only once so far (Chang et al., 2017) and not at all using total RNA-Seq data. To date, High-Throughput Sequencing (HTS) has reached sequencing depths that allow for the application of omics-based approaches in environmental studies; however, it is unclear what scales are required to allow for machine-learning-based environmental stressor predictions. There is a clear need for a comparative assessment of metagenomics, total RNA-Seq, and amplicon sequencing with respect to their ability to provide adequate taxonomic datasets for machine learning approaches.

In this study, we compare the performance of amplicon sequencing, metagenomics, and total RNA-Seq to predict environmental stressor levels based on taxonomically assigned data using machine learning. We used samples obtained from an ExStream system (Piggott et al., 2015) consisting of stream mesocosms that were exposed to fine sediment and an insecticide to investigate the impact of these aquatic key stressors on stream biodiversity and the decomposition of organic matter (Mack et al., 2022). For amplicon sequencing, we used the two marker genes ITS-2 and 16S, both with an operational taxonomic unit (OTU) clustering and an exact sequence variant (ESV) denoising method. We evaluated the markers individually as well as in combination (multi-marker approach). Stressor prediction performance (SPP) for all datasets was based on different taxonomic levels (phylum, class, order, family, genus, and species), data types (abundance, presence–absence (P–A)), feature selection (with feature selection, without feature selection), and machine learning algorithms (k-Nearest Neighbors, Linear Support Vector Classification, Logistic Ridge Regression, Logistic Lasso Regression, Multilayer Perceptron, Random Forest, Support Vector Classification, and XGBoost).

2 Materials and methods

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

2.1 Experimental setup

2.1.1 ExStream system

A detailed explanation of the ExStream system can be found in Mack et al. (2022). In summary, stream mesocosms were connected to the adjacent stream Bieber, which provided them with a constant water flow. The stream Bieber is part of the Rhine-Main-Observatory,[1] a Long-Term Ecological Research site in Germany (Haase et al., 2016; Mirtl et al., 2018). Each mesocosm was set up using substrate and organisms from the stream. A random subset of the mesocosms was exposed to either the insecticide chlorantraniliprole (Coragen, DuPont), increased fine sediment concentration, or both. Both insecticides and fine sediment are known key stressors of aquatic environments introduced into streams by agricultural runoff. The stressors were induced using a 4×2 factorial design by adding 0.2 µg/L, 2 µg/L, and 20 µg/L (acute stressor phase, 4 days) or 0.02 µg/L, 0.2 µg/L, and 2 µg/L (reduced stressor phase, 17 days) of the insecticide and 450 mL of fine sediment (<2 mm) to the mesocosms. Each possible combination of stressor levels was replicated eight times in addition to eight control mesocosms that did not receive any stressor, resulting in 64 mesocosms.

2.1.2 Assessment of microbial community compositions

The goal of the ExStream experiment was to evaluate the individual and combined effects of the applied stressors on biodiversity and organic matter decomposition in streams. To investigate organic matter decomposition, cotton strips were added to all mesocosms. Cotton strips are mainly made of cellulose, which is a major source of carbon in stream ecosystems. Therefore, analyzing the biofilm on the cotton strips allowed the analysis of the diversity of microbial communities degrading organic matter.

The experiment was divided into a colonization phase (days −21 to −1) and a stressor phase (days 0 to 21). Two cotton strips were added to each of the 64 mesocosms on day −17 (128 in total) and recovered after 28 or 35 days, respectively for more information on the phases and cotton strip addition and recovery see Mack et al. (2022). Four cotton strips were washed away during the experiment, so 124 cotton strips were recovered in total. A 2-cm-long piece of each cotton strip was cut off and transferred into a ZR BashingBead Lysis Tube (0.1 & 0.5 mm) pre-filled with 1 mL of DNA/RNAShield (Zymo Research, Freiburg, Germany) using sterile laboratory gloves, forceps, and scissors. The samples were transferred to a laboratory, stored at −20°C, and then homogenized using a bead mill homogenizer (MM 400, Retsch, Haan, Germany) at 1,800 rpm for 30 min. 300 µL of each lysate were processed for amplicon sequencing at the University of Duisburg-Essen, Germany, and the remainder of each lysate was shipped to the University of Guelph, Canada, on dry ice and processed for metagenomics and total RNA-Seq.

1 https://deims.org/9f9ba137-342d-4813-ae58-a60911c3abc1

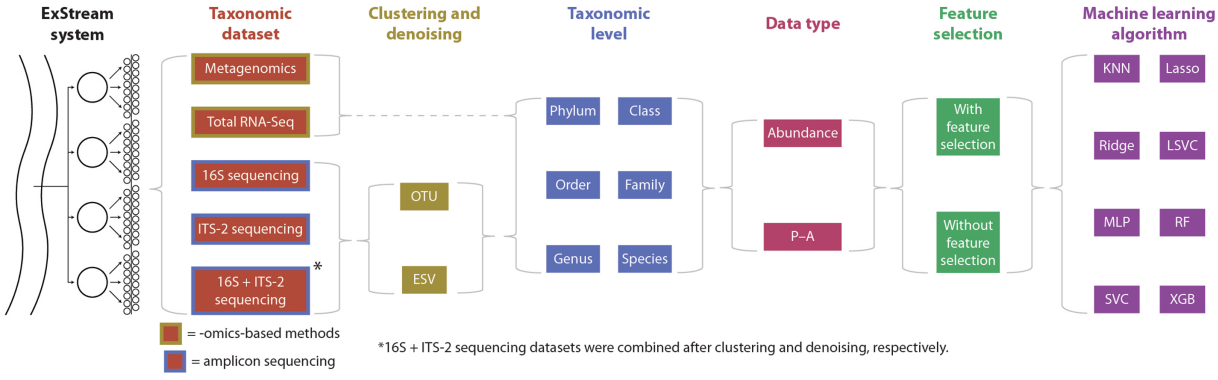

FIGURE 1
Summary of the study design. ExStream samples were processed using omics-based methods and amplicon sequencing, and HTS data were processed using two clustering methods (only applicable to amplicon sequencing), six taxonomic levels, two data types, with or without feature selection, and eight machine learning algorithms, for a total of 1,536 evaluated combinations of sequencing and data-processing methods. KNN, k-Nearest Neighbors; Lasso, Logistic Lasso Regression; Ridge, Logistic Ridge Regression; LSVC, Linear Support Vector Classification; MLP, Multilayer Perceptron; RF, Random Forest; SVC, Support Vector Classification; XGB, XGBoost.

2.2 Laboratory processing

2.2.1 Laboratory processing of amplicon sequencing

Amplicon sequencing was carried out following the workflow described by Buchner et al. (2021). All subsequent processing steps were completed on a Biomek FX^P liquid handling workstation (Beckman Coulter, Brea, CA, United States). Briefly, replication of the samples was carried out before DNA extraction by transferring 60 µL from the bead-beating tubes to deep-well plates pre-filled with 133 µL of TNES buffer (50 mM Tris, 400 mM NaCl, 100 mM EDTA, 0.5% SDS, pH 7.5) and 6 µL of Proteinase K (10 mg/mL) following incubation for 3 h at 55°C for complete lysis of the samples. DNA was extracted using a modified version of the NucleoMag Tissue kit Macherey Nagel, Düren, Germany; for modifications see Buchner et al. (2021). Extraction success was verified using a 1% agarose gel.

The PCR for the amplicon library was performed using a two-step PCR protocol following Zizka et al. (2019). Samples were amplified in a first-step PCR using the Qiagen Multiplex Plus Kit (Qiagen, Hilden, Germany) with a final concentration of 1x Multiplex Mastermix, 200 mM of each primer [515F & 806R for 16S (Caporaso et al., 2011) and ITS3-CS1 & ITS4-CS2 for ITS-2 (Frey et al., 2016)], and 1 µL of DNA, and filled up to a total volume of 10 µL with PCR-grade water. The amplification protocol was: 5 min of initial denaturation, 25 cycles of 30 s denaturation at 95°C, 90 s of annealing at 50°C for 16S and 55°C for ITS-2, and 30 s of extension at 72°C, finished by a final elongation step of 10 min at 68°C. For subsequent demultiplexing, each of the PCR plates was tagged with a unique combination of inline tags (Supplementary File S1).

The first-step PCR results were cleaned up with magnetic beads. The PCR product was mixed with clean-up buffer (2.5 M NaCl, 10 mM Tris, 1 mM EDTA, 20% PEG 8000, 0.05% Tween 20, 2% carboxylated Sera-Mag SpeedsBeads (Cytiva Life Sciences, Marlborough, MA, United States), pH 8) at a 0.8x ratio and incubated for 5 min, washed two times with wash buffer (10 mM Tris, 80% EtOH, pH 7.5) for 30 s, dried for 5 min at RT and finally eluted in 40 µL of elution buffer (10 mM Tris, pH 8.5).

During the second-step PCR, samples were amplified with a final concentration of 1x Multiplex Mastermix, 1x Coralload Loading Dye, 100 mM of each primer, and 2 µL of the first-step product. Cycling conditions were the same as in the first-step PCR except for 61°C as annealing temperature and a decreased cycle number of 20. In the second-step PCR, each of the 96 wells was individually tagged so that the combination of the in-line tag from the first-step PCR and the index-read of the second-step PCR yielded a unique combination per sample. PCR success was verified using a 1% agarose gel.

PCR products were normalized to equal concentrations with normalization buffer (same as clean-up buffer, but with only 0.1% beads) following the same protocol as the clean-up after the first step but with a ratio of 0.7x and an elution volume of 50 µL. All normalized products were pooled in the final libraries in equal parts. The libraries were concentrated using a silica-membrane spin column (Epoch Life Science, Missouri City, TX, United States) by mixing 1 volume of the library with 2 volumes of binding buffer (3 M Guanidine Hydrochloride, 90 EtOH, 10 mM Bis-Tris, pH 6) for the binding step (1 min centrifugation, 11,000 x g), 2 washing steps (30 s centrifugation, 11,000 x g) with wash buffer and a final elution (3 min incubation at RT, followed by 1 min centrifugation at 11,000 x g) with 100 µL elution buffer. Library concentrations were quantified on a Fragment Analyzer (High Sensitivity NGS Fragment Analysis Kit; Advanced Analytical, Ankeny, United States). The libraries were then sequenced using the Illumina MiSeq platform with 2 lanes for each library with a paired-end kit (V2, 2×250 bp for 16S and V3, 2×300 bp for ITS) at CeGat (Tübingen, Germany).

2.2.2 Laboratory processing of metagenomics and total RNA-Seq

DNA and total RNA were separately extracted from samples in 96-well plates using the NucleoMag DNA/RNA Water kit (D-MARK Biosciences, Toronto, Canada) that includes magnetic beads. Instead of using a magnetic plate to separate magnetic beads from buffers, we used the Magnetic Bead Extraction Replicator (V&P Scientific, San Diego, United States), which allows for the transfer of all magnetic

beads from one lysate/buffer/elution plate to another without the need to remove the supernatant from individual wells.[2] The RNA extraction protocol involved a 25-min-long rDNase incubation step to digest DNA. Since the 96-well plates were open during the entire extraction, which posed a contamination risk, we added one negative extraction control to each row of each plate by replacing lysate with pure water. All extractions were performed under a sterile hood. DNA/RNA concentrations of all extracts and all negative extraction controls were measured using a Qubit fluorometer with the dsDNA HS Assay Kit and the RNA HS ASSAY Kit, respectively (Thermo Fisher Scientific, Burlington, Canada).

DNA and RNA libraries of all samples and negative extraction controls were prepared for metagenomics and total RNA-Seq using the NEBNext Ultra II DNA Library Prep Kit for Illumina and the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina, respectively (New England Biolabs, Whitby, Canada). For RNA library preps, we did not perform mRNA enrichment or rRNA removal and instead processed the entire RNA. The RNA library prep kit has a default insert size of 200 bp, and we chose an insert size of 150–350 bp for the DNA library preps to keep insert sizes approximately consistent. After library prep, we randomly selected 8 DNA sample libraries, 3 negative DNA extraction control libraries, 7 RNA sample libraries, and 4 negative RNA extraction control libraries and sent 2.5 µL of each to the AAC Genomics Facility at the University of Guelph, Canada for analysis on an Agilent Bioanalyzer 2,100 system (Agilent Technologies, United States) to confirm successful library preps and check for contaminations in negative extraction control libraries. After consultation with the sequencing facility (Center for Applied Genomics, Hospital for Sick Children, Toronto, Canada), we cleaned up all DNA and RNA libraries following the DNA/RNA library prep kit manual to remove primer dimers and unincorporated primers.

We pooled 5 µL of each DNA and RNA library for sequencing, respectively, including negative extraction controls. We pooled equal volumes instead of equal concentrations because this pooling strategy allows for an equal relative sequencing depth per sample as opposed to an equal total sequencing depth. That way, the relative number of reads per sample mirrored the relative amount of DNA/RNA, avoiding an over- or underrepresentation of samples with higher or lower DNA/RNA amounts. Size distributions of the DNA and RNA library pools were assessed with a bioanalyzer by the sequencing facility, and the average fragment size was 386 bp for the DNA library pool and 436 bp for the RNA library pool. Both pools were paired-end (2×100 bp) sequenced in a 50:50 ratio on a single lane of a NovaSeq 6,000 SP flowcell.

2.3 Bioinformatics

2.3.1 Bioinformatics of amplicon sequencing

Raw data of the sequencing runs were delivered demultiplexed by index reads. Further demultiplexing by inline tags was done with the Python script "demultiplexer".[3] Sequences were subsequently processed with APSCALE v1.4 (Buchner et al., 2022) using default parameters. Paired-end reads were merged using vsearch v2.21.1 (Rognes et al., 2016). Primer sequences were trimmed with cutadapt v3.5 (Martin, 2011). For 16S sequencing, only sequences with a length of 252 ± 10 bp were retained, and for ITS-2 sequencing, only sequences with a length ranging from 240 to 460 bp were retained. Only sequences with an expected error of 1 passed quality filtering. Reads were dereplicated and singletons were removed. For OTU generation, sequence clustering was performed with a similarity threshold of 97%, and for ESV generation, denoising was carried out with an alpha value of 2 and a minimum size of 8 as implemented in vsearch. Before taxonomic assignment, the resulting OTU and ESV tables were filtered for potentially biased sequences using the LULU algorithm (Frøslev et al., 2017) implemented in APSCALE.

Subsequently, only OTUs and ESVs found in both replicates of the same sample were summed up for all samples. After this initial data filtering, reads still left in the negative controls were subtracted from OTUs or ESVs, respectively, to generate final OTU and ESV tables. Taxonomic assignment was performed using DADA2 with default parameters in combination with the database SILVA 138.1 designed for DADA2 (McLaren and Callahan, 2021) for 16S sequences and the database UNITE (Abarenkov et al., 2021) for ITS-2 sequences, respectively.

2.3.2 Bioinformatics of metagenomics and total RNA-Seq

In an earlier study, we investigated 672 combinations of bioinformatic tools to identify the best-performing combination to process and taxonomically annotate microbial mock community datasets (Hempel et al., 2022). Based on these results, we processed both metagenomics and total RNA-Seq data as follows: we used Trimmomatic v0.39 (Bolger et al., 2014) to trim the leading and trailing low-quality nucleotides of each read by cutting reads if the average quality of nucleotides in a sliding window of size 4 was below a PHRED score of 20. After trimming, we excluded reads shorter than 25 nucleotides and error-corrected reads using the error-correction module of the assembler SPAdes v3.14.1 (Bankevich et al., 2012). Then we assembled the reads into scaffolds using MEGAHIT v1.2.9 (Li et al., 2015) with the parameter 'presets' set to 'meta-large' to adjust k-mer sizes for the assembly of large and complex metagenomes. All other parameters were set to default. Subsequently, we mapped reads to assembled scaffolds to determine the abundance of each scaffold using BWA v0.7.17 (Li and Durbin, 2009) with default parameters. We processed mapped reads using the function *coverage* of samtools v1.10 (Li et al., 2009) to obtain the mean per-base coverage for each scaffold. For taxonomic annotation, we used the SILVA132_NR99 SSU and LSU reference databases (Quast et al., 2013) in combination with kraken2 v2.1.1 (Wood et al., 2019) using default parameters. The setup of the kraken2 database for SILVA required manual adaptations, which are described in the Supplementary material. All code utilized is available on GitHub.[4]

2 For the modified protocol, see dx.doi.org/10.17504/protocols.io.bp2l69n2dlqe/v1

3 v1.1.0, https://github.com/DominikBuchner/demultiplexer

4 https://github.com/hempelc/metagenomics-vs-totalRNASeq

2.4 Pre-processing of taxonomic data

The data were further processed in Python v3.7.9 (Van Rossum and Drake, 2009). The full code is available on GitHub[5] and involves the modules Pandas v1.3.5 (Reback et al., 2021) and NumPy v1.21.3 (Harris et al., 2020). We trained and evaluated machine learning models based on phylum, class, order, family, genus, and species to assess differences in Stressor prediction performance (SPP) among taxonomic levels. Because both metagenomics and total RNA-Seq datasets consisted of mean per-base coverage while amplicon sequencing datasets consisted of absolute read counts, we employed two different approaches to determine taxa abundances for each taxonomic level. When aggregating metagenomic and total RNA-Seq taxonomic datasets for each level separately, we adjusted taxa abundances for sequencing depth and scaffold length. For that, we selected all scaffolds assigned to each detected taxon and determined each taxon's absolute abundance as follows:

$$perBcov_{taxon} = \frac{covered\ bases\ across\ scaffolds}{total\ bases\ across\ scaffolds} = \frac{\sum_1^{scaf} perBcov_{scaf} \times len_{scaf}}{\sum_1^{scaf} len_{scaf}},$$

where $perBcov_{taxon}$ represents the per-base coverage of a taxon, *scaf* represents the number of scaffolds assigned to a taxon, and $perbcov_{scaf}$ and len_{scaf} represent the per-base coverage and length of each scaffold. We then converted absolute abundances into relative abundances. This process is similar to that for abundance estimation of binned scaffolds (Parks et al., 2015).

When aggregating abundances based on amplicon sequencing data for each taxonomic level separately, we determined absolute taxa abundances as the cumulative read count of each detected taxon and converted absolute abundances into relative abundances.

For metagenomics and total RNA-Seq samples, negative extraction controls were subtracted from samples that were co-extracted with the controls. We converted relative abundances into absolute abundances by multiplying relative abundances by the number of reads per sample, summarized the absolute abundances of taxa among all negative extraction controls per plate, and subtracted the cumulative absolute abundance of each taxon detected within controls from the actual samples of the same plate. Afterwards, we reverted absolute abundances back into relative abundances.

We then excluded the taxonomic entry *NA* from all datasets, which represented the relative abundance of sequences that could not be taxonomically annotated, likely due to missing references in databases or sequencing and data-processing errors. Next, we readjusted the relative abundances of all other taxa. In some datasets, some samples consisted only of sequences that could not be taxonomically annotated, meaning that they had a cumulative relative abundance of zero after excluding the *NA* entry. These samples were considered to have failed, and we excluded them from all datasets to ensure that all datasets contained the same samples, which ultimately resulted in 121 samples per dataset.

To assess differences in SPP among data types, we evaluated abundance and P–A data. For P–A data, we set all relative abundances above 0 to 1 (0 = absent, 1 = present). For abundance data, we followed the appropriate steps for analyzing compositional data, as pointed out by Gloor et al. (2017). Therefore, we first applied simple multiplicative replacement to replace zeros among all relative abundances using the function *multiplicative_replacement* of the Python module scikit-bio v0.5.6 (The Scikit-Bio Development Team, 2020). The function replaces zeros with a small positive value δ, which is based on the number of taxa while ensuring that the compositions still add up to 1. Then, we applied a centered log-ratio (clr) transformation using the function *clr* of scikit-bio, which captures the relationships between taxa and makes the data symmetric and linearly related. Since feature standardization is required by some machine learning algorithms, we further standardized taxa abundances using the function *StandardScaler* of the Python module scikit-learn v1.1.1 (Pedregosa et al., 2011).

To include a multi-marker approach using both the ITS-2 and 16S marker genes in the evaluations, we combined the generated 16S and ITS-2 datasets by concatenating them using the clustering or denoising method (OTUs or ESVs). This resulted in eight taxonomic datasets that were evaluated (ITS-2 amplicon sequencing clustered into OTUs (ITS-2 OTU) or denoised into ESVs (ITS-2 ESV), 16S amplicon sequencing clustered into OTUs (16S OTU) or denoised into ESVs (16S ESV), multi-marker approach clustered into OTUs (16S + ITS-2 OTU) or denoised into ESVs (16S + ITS-2 ESV), metagenomics, and total RNA-Seq).

2.5 Biodiversity analysis

To analyze the biodiversity detected per taxonomic dataset, we grouped detected taxa using NCBI GenBank taxonomy. We determined the total number of detected taxa per taxonomic dataset, the number of unique taxa detected within only one taxonomic dataset, and the number of overlapping taxa between taxonomic datasets at the phylum, genus, and species level. For that we translated all phyla, genus, and species names within 16S and ITS-2 datasets into NCBI taxonomy to match names across all datasets and utilized reference databases. Specifically, we tested each name for matches with names in the scientific or non-scientific NCBI taxonomy,[6] and if a match was found, the name was translated into the scientific NCBI name. If no match was found, we manually checked if the respective name was available on NCBI under a different scientific or non-scientific name, and if so, the alternative scientific name was used. Otherwise, the name was not available on NCBI and was used without translation. After translation, taxa containing the terms "candidatus," "candidate," or "[candida]" were removed. Then, the number of overlapping taxa between taxonomic datasets was determined as the number of matches between the respective taxa within each taxonomic dataset, and the number of

5 https://github.com/hempelc/exstream-metagenomics-totalrnaseq-ml

6 NCBI taxonomy file *names.dmp*, available through the NCBI archive as part of *taxdmp.zip*, https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/

taxa unique to one taxonomic dataset was determined by subtracting the number of overlapping taxa from the total number of detected taxa.

2.6 Machine learning

2.6.1 Data preprocessing

Taxon abundances/P–A represented independent features, and we defined the dependent feature as the combinations of applied insecticide level (none, low, medium, high) and fine sediment addition (normal fine sediment concentration, increased fine sediment concentration) for each sample, resulting in eight classes that were predicted by the machine learning algorithms. Since correlated independent features add noise, we removed them by applying the SULOV (Searching for Uncorrelated List of Variables) algorithm using the function *FE_remove_variables_using_SULOV_method* of the Python module featurewiz v0.1.55,[7] which identifies all pairs of highly correlated independent features (features with a Pearson correlation coefficient of >0.7 or < −0.7 by default), determines their Mutual Information Score (MIS) to the dependent feature, and keeps the independent feature with the highest MIS for each highly correlated feature pair.

2.6.2 Test-train splitting and feature selection

Each ExStream mesocosm was sampled at two time points as part of the cotton strip assay, which meant that samples consisted of highly related paired samples, i.e., two samples of the same mesocosm. When splitting the data sets into train and test sets, we ensured that paired samples were assigned to the same training and test sets to avoid data leakage between the sets.

Initially, we applied a 90:10 train-test split to the datasets (109 train samples, 12 test samples) and performed training and testing without repetition, but due to large discrepancies between train and test scores, we changed the train-test split ratio to 80:20 (97 train samples, 24 test samples) and repeated both training and testing splits three times in total. During each repetition, we randomly selected 12 pairs (24 samples) of highly related samples for the test dataset and trained and tested all models across all datasets with the same randomly selected 12 sample pairs per repetition.

For feature selection, we used Recursive Feature Elimination to select the 20 most important features using the function *RFE* from scikit-learn with a *DecisionTreeClassifier* as the estimator.

2.6.3 Model selection, training, and testing

It is generally recommended to test multiple machine learning algorithms (Greener et al., 2022), which is why we selected eight machine learning algorithms to predict stressor classes: k-Nearest Neighbors (KNN), Linear Support Vector Classification (LSVC), Logistic Ridge Regression (Ridge), Logistic Lasso Regression (Lasso), Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Classification (SVC), and XGBoost (XGB). For thorough descriptions of these algorithms in a biological context see Greener et al. (2022) and Ghannam and Techtmann (2021).

All algorithms, except XGBoost, are available in scikit-learn. To run the XGBoost algorithm, we used the Python module xgboost v1.6.1 (Chen and Guestrin, 2016), which is compatible with scikit-learn. To optimize hyperparameters while avoiding overfitting, we performed Bayesian hyperparameter optimization with 10-fold cross-validation using the function *BayesSearchCV* of the Python module scikit-optimize v0.9.0.[8] The function is compatible with scikit-learn and builds a performance probability model for given hyperparameters, which is used to select the most promising hyperparameters through iterative performance evaluations. While not every possible hyperparameter combination is tested that way, this approach provides a good trade-off between optimization results and runtime. Model prediction performance was evaluated using the Matthews Correlation Coefficient (MCC), which ranges from −1 to 1, where 1 means perfect predictions/performance, 0 means prediction performance as good as random guessing, and − 1 means all predictions are wrong, and increments between −1 and 1 can be interpreted in the same way as increments of the Pearson correlation coefficient. All hyperparameters tested can be found in the publicly available code[9] and Supplementary File S2. The optimized hyperparameters were then used to train models on the entire training dataset, and model performances to predict classes of the testing dataset were evaluated using the MCC. During training on the entire dataset, learning curves were generated using the *learning_curve* function from scikit-learn. This process was repeated three times, as described above, and the mean average and standard deviation (SD) of the training and test MCC scores across the three repetitions were determined.

We tested each possible combination of taxonomic datasets (ITS-2, 16S, 16S + ITS-2, metagenomics, and total RNA-Seq), clustering or denoising methods (OTU, ESV; only applicable to amplicon sequencing data), taxonomic levels (phylum, class, order, family, genus, and species), data types (abundance, P–A), feature selection (with feature selection, without feature selection), and classification algorithms (KNN, Lasso, LSVC, Ridge, MLP, RF, SVC, and XGB), resulting in a total of 1,536 evaluated combinations.

2.7 Statistical analysis

We quantified the impact of sequencing types, taxonomic levels, data types, feature selection, and machine learning algorithms on SPP. For that, we converted all sequencing and data-processing methods into binary dummy variables and tested for significant correlations ($p \leq 0.05$) between each sequencing and data-processing method and the test MCC by calculating Spearman's rank correlation coefficient using the *spearmanr* function of the Python module SciPy v1.7.1 (Virtanen et al., 2020). Additionally, we performed the same test for each sequencing type separately.

7 https://github.com/AutoViML/featurewiz

8 https://github.com/scikit-optimize/scikit-optimize

9 https://github.com/hempelc/exstream-metagenomics-totalrnaseq-ml

3 Results

3.1 High-throughput sequencing results

We obtained 248,707,817 paired-end reads from metagenomics [mean average per sample: 2 M reads, standard deviation (SD): 2.4 M reads], 206,096,238 from total RNA-Seq (mean average per sample: 1.7 M reads, SD: 2.6 M reads), 21,719,985 reads from 16S sequencing (mean average per sample: 152 k reads, SD: 27 k reads), and 27,033,469 reads from ITS-2 sequencing (mean average per sample: 214 k reads, SD: 41 k reads; Supplementary Figure S1; Bioproject number: PRJNA903104, SRA accession numbers: SRR22331748–SRR22332597). The SD of the mean average number of metagenomics and total RNA-Seq reads per sample was very high because we normalized metagenomics and total RNA-Seq libraries based on volume during library preparation so that the relative number of reads per sample mirrored the relative amount of DNA/RNA. This avoided an over-or underrepresentation of samples with higher or lower amounts of DNA/RNA but also led to substantial variations in the number of reads per metagenomics/total RNA-Seq library (Supplementary Figure S1).

3.2 Biodiversity analysis

There were no taxa overlaps between ITS-2 and 16S sequencing at the phylum, genus, and species level (Figure 2, for exact numbers, see Supplementary File S3), while either method had overlapping taxa with both metagenomics and total RNA-Seq. Metagenomics and total RNA-Seq shared more taxa with each other than with ITS-2 or 16S sequencing. Metagenomics detected by far the most phyla (95), genera, (2488), and species (3,522), and the number of genera and species detected using metagenomics was much higher relative to that of other taxonomic datasets than the number of detected phyla. For total RNA-Seq, the number of detected phyla (76) was more than three times as high as that of ITS-2 (OTU: 23, ESV: 20) and 16S sequencing (OTU: 25, ESV: 24), the number of detected genera (903) was 1.3–2.9 times as high as that of ITS-2 sequencing (OTU: 678, ESV: 491) and 16S sequencing (OTU: 315, ESV: 363), and the number of detected species (892) was 1.3–1.8 times as high as that of ITS-2 sequencing (OTU: 673, ESV: 506) and much higher than that of 16S sequencing (OTU: 55, ESV: 114). 16S sequencing detected almost the same number of phyla as ITS-2 sequencing but by far the lowest number of genera and species among all taxonomic datasets. In terms of taxa unique to one taxonomic dataset, metagenomics detected by far more unique phyla (19), genera (1,399), and species (2660) than all other all other taxonomic datasets combined. Within ITS-2 and 16S sequencing, OTU clustering and ESV denoising resulted in different numbers of detected taxa, specifically for ITS-2 sequencing at genus level (OTU: 678, ESV: 491) and for 16S sequencing on the species level (OTU: 55, ESV: 114). 16S sequencing detected much less taxa at species level than at genus level. In terms of the distribution of taxonomic groups, 16S sequencing recovered almost exclusively bacterial taxa, while ITS-2 sequencing recovered not only taxa in the group "plants and fungi" but also invertebrate taxa. Omics-based methods recovered taxa across all groups, and they detected more bacterial taxa than 16S sequencing at all three taxonomic levels. At genus and species level, bacterial taxa represented most detected taxa.

3.3 Impact of taxonomic datasets and data-processing methods on SPP

SPP varied substantially across tested combinations of taxonomic datasets, clustering or denoising methods, taxonomic levels, machine learning algorithms, and feature selection (Figure 3; since data types had no significant impact on SPP (see Figures 4, 5), only P–A-based SPPs are shown). MCC values ranged from below 0 (prediction SPP worse than random guessing) to 0.45 (moderate to good SPP). Feature selection overall improved SPP. ITS-2 sequencing and omics-based

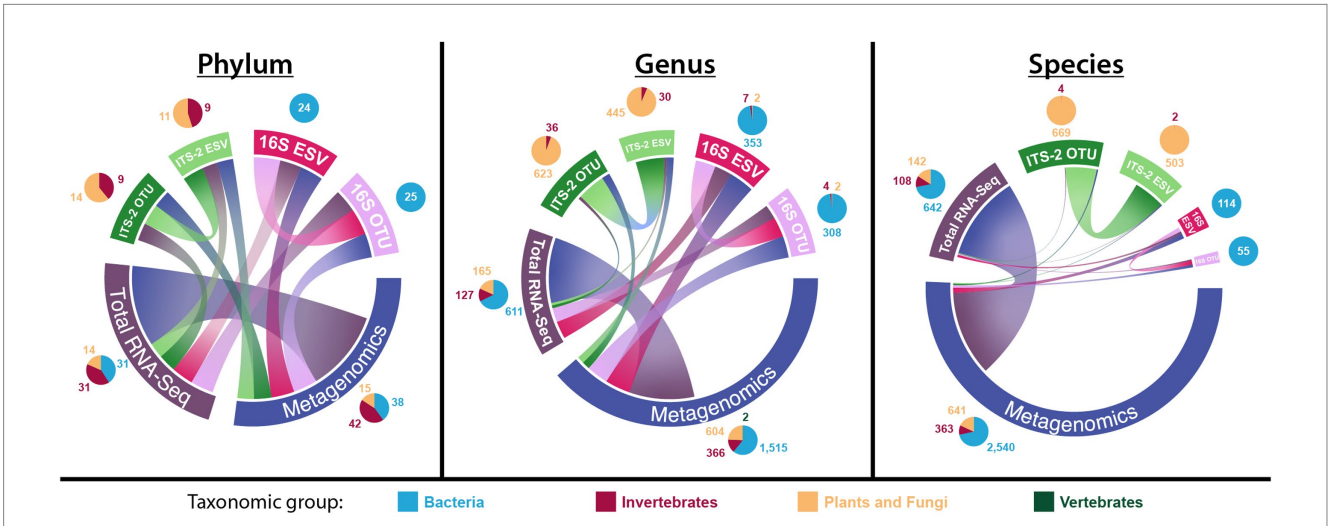

FIGURE 2
Number of total, unique, and overlapping taxa for each taxonomic dataset on the phylum, genus, and species level (chord diagrams), as well as the distribution of taxonomic groups within each taxonomic dataset (pie charts). In the chord diagrams, the size of the outer bars represents the total number of detected taxa, the size of the connections between taxonomic datasets represents the number of overlapping taxa, and the fraction of outer bars with no connection to other taxonomic datasets represents the number of unique taxa detected only in that taxonomic dataset.

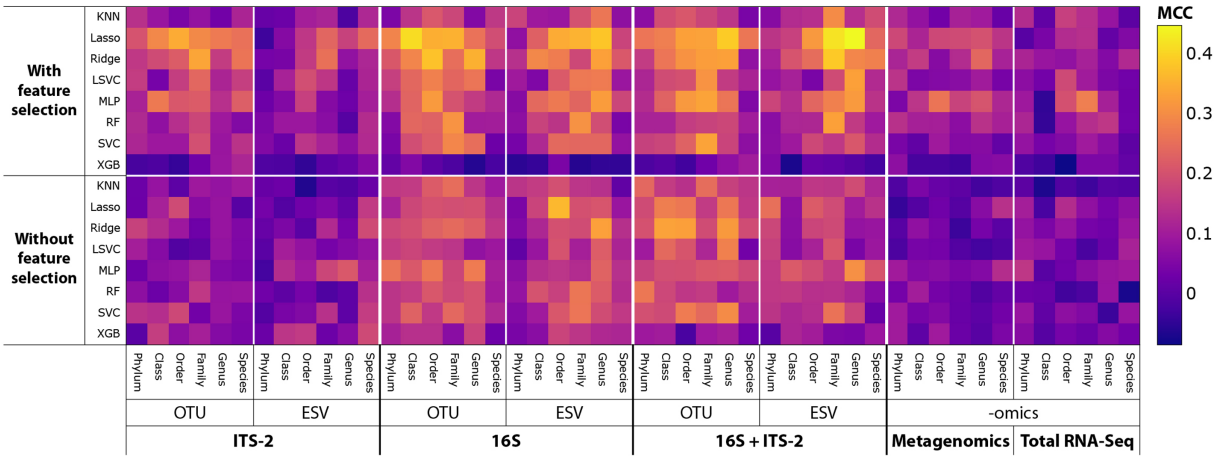

FIGURE 3
MCC as a proxy for SPP across all combinations of sequencing and data-processing methods tested. Since data types had no significant impact on SPP (see Figures 4, 5), only P–A-based SPPs are shown.

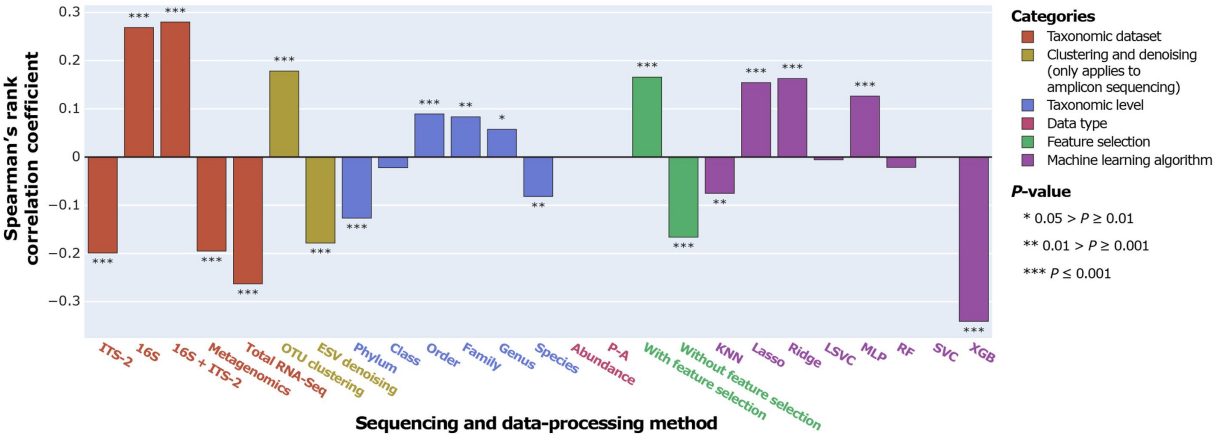

FIGURE 4
Correlation between MCC as a proxy for SPP and sequencing and data-processing methods.

methods performed poorly overall, except for some combinations of ITS-2 sequencing with OTU clustering, whereas 16S sequencing and the multi-marker approach of combined 16S and ITS-2 markers performed better overall. The highest MCC of 0.45 was found for the following combination: 16S + ITS-2 sequencing, ESV denoising, genus level, P–A data, Lasso algorithm, with feature selection. For this combination, the learning curves generated during each training repetition indicated that the model was overfitted, meaning that more data, i.e., more samples would have likely further increased SPP (Supplementary Figure S2).

Overall, ITS-2 sequencing, metagenomics, and total RNA-Seq significantly negatively correlated with SPP, and 16S sequencing and combined 16S + ITS-2 markers significantly positively correlated with SPP (Figure 4). For amplicon sequencing, OTU clustering significantly increased SPP while ESV denoising significantly decreased SPP. Performance increased with increasing taxonomic resolution up to the order level and decreased at higher levels. Data types did not significantly correlate with SPP. Feature selection significantly increased SPP. SPPs varied between machine learning algorithms, with XGB performing by far the worst and Lasso and Ridge, which are both based on logistic regression, performing the best, followed by MLP.

The impact of data-processing methods on SPP varied between individual taxonomic datasets (Figure 5). For ITS-2 ESV, the species level was significantly positively correlated with SPP, which contrasted with all other taxonomic datasets. For metagenomics, no taxonomic level significantly correlated with SPP. Data types did not significantly correlate with SPP in any taxonomic dataset. Feature selection had the strongest impact on metagenomics and no significant impact on 16S OTU/ESV and 16S + ITS-2 OTU. Across all taxonomic datasets, XGB performed poorly. Lasso and Ridge performed significantly well for all taxonomic datasets except metagenomics, total RNA-Seq, and ITS-2 ESV. Overall, the impact of data-processing methods was similar between 16S OTU/ESV, 16S + ITS-2 OTU/ESV, and ITS-2 OTU and differed between metagenomics, total RNA-Seq, and ITS-2 ESV.

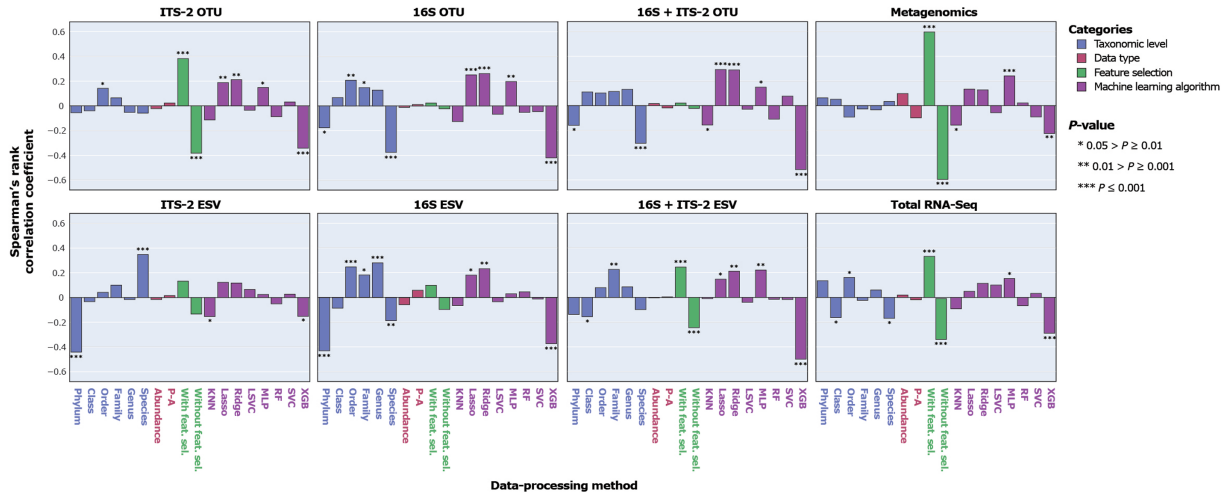

FIGURE 5
Correlation between MCC as a proxy for SPP and data-processing methods for individual taxonomic datasets.

4 Discussion

4.1 Biodiversity analysis

The number of total, unique, and overlapping taxa varied substantially between taxonomic datasets. ITS-2 and 16S sequencing had no taxa overlap, confirming that both markers were group-specific; however, while 16S sequencing was almost exclusively specific to bacteria, ITS-2 sequencing detected not only taxa belonging to the NCBI division "plants and fungi" but also invertebrate taxa, indicating that the applied ITS-2 primers were not specific to fungi. Metagenomics and total RNA-Seq had overlapping taxa with ITS-2 and 16S sequencing but also detected a high number of taxa that the latter did not detect, and both methods detected bacterial, invertebrate, plant and fungal taxa, confirming that omics-based methods can recover groups across the tree of life, which is considered a major advantage over amplicon sequencing (Shakya et al., 2013; Brumfield et al., 2020; Obiol et al., 2020). Many taxa found with total RNA-Seq were also found with metagenomics, but the latter also found an extremely high number of unique taxa. However, at genus and species level, ITS-2 sequencing detected a high number of unique taxa as well. These taxa were not recovered by omics-based methods, potentially because we only utilized SSU and LSU references for taxonomic annotation of omics-based sequences, or because ITS-2 sequencing has a higher taxonomic resolution within fungi than omics-based methods at our utilized sequencing depth. In contrast, omics-based methods found much more bacterial species, genera, and even phyla than 16S sequencing. While metagenomics can identify bacterial taxa at the species or even strain level given sufficient sequencing depth, 16S sequencing is often limited to bacterial genus level identifications (Knight et al., 2018), which could explain why 16S sequencing detected fewer bacterial species than genera and fewer bacterial species than omics-based methods. However, the fact that omics-based methods also detected much more bacterial taxa at genus and even phylum level shows that either the taxonomic resolution of omics-based methods outperformed that of 16S sequencing for bacteria or that these methods detected a high number of false-positive bacterial taxa. There is no clear consensus in the literature as to which of those methods detect more taxa, with some studies showing that amplicon sequencing detects more taxa than omics-based methods (Stat et al., 2017; Tessler et al., 2017), while others show that both methods detect equal amounts of taxa (Chan et al., 2015; Obiol et al., 2020) or that omics-based methods outperform amplicon sequencing in terms of biodiversity coverage (Shakya et al., 2013; Laudadio et al., 2018; Yan et al., 2018; Brumfield et al., 2020). Biodiversity coverage also depends on how well an environment is represented in reference databases, and for less-studied environments that are poorly represented in reference databases, it is possible that the majority of omics-based sequences cannot be taxonomically annotated, resulting in low overall taxonomic resolution (Stat et al., 2017). Our results support both hypotheses: (1) omics-based methods detect more taxa overall, and (2) amplicon sequencing detects more taxa within target groups, at least for fungi, which aligns with the advantages and disadvantages of either approach. In theory, all taxa detected with amplicon sequencing should also have been detected with omics-based methods, but our results indicate that sequencing depth for omics-based methods must be increased substantially to be able to detect the same taxa. Tools and databases that incorporate references from more taxonomic markers to identify omics-based sequences should also be further explored. However, given continuous technological advancements in HTS capacities, sufficient sequencing depths should become more affordable, and in combination with the steady growth of reference databases, we expect omics-based methods to unilaterally detect more taxa than amplicon sequencing at equal or higher taxonomic resolution in the future.

4.2 Impact of sequencing methods on SPP

SPP varied substantially among taxonomic datasets. 16S sequencing was the only standalone method positively correlated with SPP, and combining 16S with ITS-2 sequencing data slightly improved SPP. We expected omics-based methods to outperform amplicon

sequencing because the former are not group-specific and can cover biodiversity across the tree of life, providing a more complete picture of microbial communities; however, the opposite was the case, indicating that while omics-based methods did detect more taxa, they also missed crucial taxa, detected taxa without correlation to stressors, and/or generated more noise, which decreased SPP. This was further supported by the fact that the SPP of metagenomics, which detected the highest number of taxa, improved substantially under feature selection, i.e., the exclusion of all but the 20 most relevant taxa for model performance. However, even with feature selection, metagenomics still showed poor overall SPP, indicating that the feature-selected taxa did not include crucial taxa, did not correlate with stressors, or were poorly represented. This could be a result of insufficient sequencing depth, possibly causing insufficient recovery of taxa, or of the utilized reference database (SILVA), which only contains SSU and LSU sequences and no other commonly used markers or whole genome sequences, decreasing the likelihood of finding a taxonomic match among omics-based sequences.

Typical metagenomics experiments aim to generate between 1 and 10 Gb of metagenomic data per sample (Quince et al., 2017) while we generated on average 0.2 Gb metagenomic data per sample, which is one to two magnitudes lower. Increasing the sequencing depth of omics-based methods to ensure that taxa with high bioindication potential are sufficiently represented might increase SPP but is currently also related to substantially higher costs. In previous studies, we showed that total RNA-Seq outperformed metagenomics in identifying a microbial community and reconstructing SSU rRNA sequences (Hempel et al., 2022, 2023) at lower sequencing depth and, therefore, costs, likely due to higher SSU rRNA sequence yield when using total RNA-Seq. Therefore, for the present study, we expected that total RNA-Seq would have a higher SPP than metagenomics at comparably low sequencing depth (on average 0.17 Gb total RNA-Seq data per sample). However, total RNA-Seq performed even worse, indicating that even the sequencing depth of total RNA-Seq was too low.

The poor performance of metagenomics could also be related to the fact that only SSU and LSU reference sequences were used for taxonomic annotation instead of all available markers or whole genome sequences to utilize all available metagenomic information. In the present study, we compared metagenomics and total RNA-Seq explicitly due to the aforementioned advantages of total RNA-Seq in regard to SSU and LSU rRNA coverage. Therefore, testing databases and tools that incorporate more markers or whole genome sequences for taxonomic annotation, such as MetaPhlAn (Blanco-Míguez et al., 2023) or the NCBI Genbank database, was out of scope for this study; however, due to the poor performance of both omics-based methods, these options should be further explored in similar future studies.

Almost all studies that utilize machine learning for taxonomically assigned HTS data in an ecological context involve amplicon sequencing (Smith et al., 2015; Cordier et al., 2017, 2018; Gerhard and Gunsch, 2019; Frühe et al., 2020; Hermans et al., 2020; Dully et al., 2021), and to our knowledge, there is only one study that involves metagenomics in that context (Chang et al., 2017) and none that compare amplicon sequencing with omics-based methods. However, in a medical context, Marcos-Zambrano et al. (2021) provide a thorough overview of human microbiome studies that utilize machine learning for HTS data. While they list seven studies that applied machine learning to both amplicon sequencing and metagenomics data, only one of them compared the performance of both sequencing methods based on community composition (Douglas et al., 2018), showing that amplicon sequencing outperformed metagenomics in classifying patients and the state of Crohn's disease while metagenomics outperformed amplicon sequencing in classifying treatment response. These results further demonstrate that SPP is dependent on the environmental variables investigated. Multiple other medical studies utilizing machine learning for disease predictions based on metagenomics community compositions show good SPP for predicting colorectal cancer, inflammatory bowel disease, diabetes, rheumatoid arthritis, and liver cirrhosis (Hacilar et al., 2018; Wu et al., 2018; Ai et al., 2019). These studies clearly show the potential of omics-based methods for medical applications, and further omics-based ecological research with sufficient sequencing depth is required to show if the methods hold the same potential for environmental stressor predictions.

4.3 Impact of data-processing methods on SPP

Data-processing methods had a substantial impact on SPP, and based on the utilized methods, SPP could range from low to high within one taxonomic dataset.

4.3.1 Impact of clustering and denoising methods on SPP

For amplicon sequencing data, OTU clustering significantly improved SPP while ESV denoising significantly decreased SPP. This observation is in contrast to the emerging recommendation to denoise amplicon sequences into ESVs (Callahan et al., 2017; Knight et al., 2018). Studies comparing OTU clustering and ESV denoising approaches did not yet reach a consensus, showing that either both approaches lead to similar results (Glassman and Martiny, 2018; Vera-Gargallo et al., 2019; Kang et al., 2021), ESV denoising outperforms OTU clustering (Caruso et al., 2019; Tapolczai et al., 2019; Joos et al., 2020), or vice versa (Roy et al., 2019; Tedersoo et al., 2022). Our results support the latter, although more similar studies are required to determine if clustering or denoising is more appropriate for machine-learning-based environmental predictions using microbial communities.

4.3.2 Impact of taxonomic levels on SPP

In general, a higher taxonomic resolution provides a better picture of microbial communities, but our results show that the species level correlated worse with SPP than genus, family, order, and even class levels. For ITS-2 sequencing and omics-based methods, the high number of detected taxa at the species level might have added more noise than value to the data. This is indicated by the significantly positive impact of feature selection on SPP, i.e., the limitation of the number of included taxa. However, for 16S sequencing, feature selection had no impact on SPP while the species level still negatively correlated with SPP. This result may be related to the number of sequences that could not be assigned to the species level and were consequently dropped. The lower the taxonomic level considered, the harder it is to annotate taxonomy due to the lack of reference sequences in databases, and the more sequences are dropped from the downstream analysis. In microbiome amplicon sequencing studies,

the taxonomic resolution is usually limited to the genus level due to the difficulty in designing primers that resolve microbial communities at the species level (Knight et al., 2018). Metagenomics allows for taxonomic resolutions at the species level or even strain level, but this requires sufficient sequencing depth (Knight et al., 2018). Dropping sequences from the analysis is equivalent to a loss of information, which could have decreased SPP at the species level. It is also possible that correlations between taxa and environmental variables are higher at lower taxonomic levels because lower taxonomic groups can be overall ecologically coherent, i.e., share similar physiologies, while higher taxonomic groups can be ecologically incoherent and have very different physiologies (Philippot et al., 2010; Choe et al., 2021; Auladell et al., 2022). Once reference databases have been extensively expanded and most sequences can be taxonomically annotated, it will be possible to determine if the lack of reference sequences or ecological incoherency of species explains lower SPP at the species level.

4.3.3 Impact of data types on SPP

We were surprised that the data types (abundance/P–A) did not have an impact on SPP, given that many studies focus on methods to improve abundance estimates from HTS data (Dillies et al., 2013; Gloor et al., 2017; Weiss et al., 2017; Pereira et al., 2018). The difference in abundance and P–A data lies in the weight of the taxa; in P–A data, abundant and rare taxa are weighted equally, making the data more sensitive to noise but also to subtle differences in community composition. Using simulated data, Koh et al. (2019) demonstrated that P–A data is more powerful when taxa associated with an environmental variable are rare while abundance data is more powerful when those taxa are abundant. However, a large-scale morphological study on benthic invertebrates showed that ecological status classifications based on abundance and P–A data showed only minor variations (Buchner et al., 2019). In a microbial context, multiple HTS studies showed similar correlations of both abundance and P–A data with environmental variables (Muletz Wolz et al., 2018; Knowles et al., 2019; Farinella et al., 2022), while some studies showed that correlations differed between data types (Kask et al., 2020; Tavalire et al., 2021). These results indicate that the impact of data types might depend on the studied environmental variables, but if further research shows that both data types have similar predictive power for environmental assessments, as our results suggest, then P–A data could be used exclusively in future environmental assessment studies. This would avoid the rather complex and partially disagreeing statistical methods required when working with compositional data, i.e., HTS abundance data (Dillies et al., 2013; Gloor et al., 2017; Weiss et al., 2017; Pereira et al., 2018). Furthermore, if abundance and P–A data generate similar results, then the often-stated advantage of metagenomics to generate abundance data free from target PCR bias (Knight et al., 2018; Khachatryan et al., 2020) would become irrelevant, which would decrease the value of omics-based approaches in comparison to amplicon sequencing.

4.3.4 Impact of feature selection on SPP

Feature selection can be applied to microbial data to remove noninformative, noisy, or redundant features (Ghannam and Techtmann, 2021). This is generally recommended because the high number of observed features can increase the risk of overfitting, which is described as the "curse of dimensionality" (Oudah and Henschel, 2018). However, feature selection goes against the proposed idea that a more holistic picture of environmental microbial communities is beneficial for predicting environmental variables, as it reduces the number of taxa included in prediction models. Our results suggest that feature selection improves SPP overall and especially for metagenomics, while the SPP of 16S sequencing was not impacted by feature selection. This indicates that the increased biodiversity coverage of omics-based methods might in fact not be beneficial for machine learning predictions and that datasets covering a lower number of taxa, as generated by amplicon sequencing, might result in more accurate and precise predictions. It should be noted, though, that ITS-2 sequencing detected approximately as many species as total RNA-Seq, and feature selection did increase the SPP of ITS-2 sequencing, showing that amplicon sequencing can also be significantly impacted by feature selection. Furthermore, the sequencing depth of metagenomics and total RNA-Seq in our study was very low, which could have influenced the impact of feature selection. If similar studies with a sufficient sequencing depth come to the same conclusion that omics-based methods in fact detect too many taxa for accurate and precise environmental assessments and require feature selection, then this would strongly tip the balance in favor of amplicon sequencing.

4.3.5 Impact of machine learning algorithms on SPP

Machine learning algorithms had a substantial impact on SPP, and even when applying two different algorithms to the same data set, the resulting MCC could range from 0.38 to −0.05. This illustrates the importance of testing multiple machine learning algorithms, which is recommended in general (Greener et al., 2022). One of the most commonly applied machine learning classification algorithms for HTS data is RF (Smith et al., 2015; Frühe et al., 2020; Hermans et al., 2020; Lanzén et al., 2020; Dully et al., 2021; Ghannam and Techtmann, 2021; Marcos-Zambrano et al., 2021), which reveals which feature contributed most to a prediction. Other popular algorithms are XGB, Support Vector Machines (which include SVC and LSVC), Logistic Regression, and KNN (Ghannam and Techtmann, 2021; Marcos-Zambrano et al., 2021; Greener et al., 2022). However, among those algorithms, RF and (L)SVC did not significantly correlate with SPP in our study, while XGB and KNN significantly negatively correlated with SPP and only logistic regression, specifically Lasso and Ridge, significantly positively correlated with SPP. Linear algorithms have the lowest flexibility among all popular machine learning algorithms, since they assume only linear relationships, and while other algorithms can assume non-linear relationships, which increases their flexibility and is often considered beneficial for the analysis of large and complex data, this was not the case for our study. In contrast, MLP, which represents a simple neural network (NN) with the highest flexibility among all algorithms tested in our study, performed overall the best after Lasso and Ridge and specifically the best for omics-based methods that generated the largest datasets. NNs are currently among the most powerful machine learning algorithms for the analysis of extremely large data, and their impact is so significant that an entirely new field of research emerged around NNs, called deep learning (Greener et al., 2022). To unfold their potential, NNs require large amounts of samples that usually go beyond the number of samples generated in a single

biological study. However, thousands of sampling sites are monitored for routine environmental assessments, and once the broad application of omics-based methods becomes more affordable, it will be interesting to see if NNs are required for good SPP based on omics data or if less complex machine learning algorithms will be sufficient or even more appropriate.

Overall, our study shows that data-processing methods should be chosen carefully since they can have a high impact on SPP and that methods resulting in the single best SPP are not necessarily the most appropriate overall. Therefore, we conclude that it is advisable to explore multiple sequencing and, in particular, data-processing methods to maximize prediction performance.

4.4 Perspectives for ecological assessments

The highest MCC, i.e., the best SPP observed in our study was 0.45, indicating moderate to good performance. This is promising, but stressor predictions must be more accurate and precise to reach the standard for applied ecological assessments. However, while the stressors tested in our study (insecticide and increased fine sediment deposition) have direct negative effects on typical indicator organisms (e.g., benthic macroinvertebrates), little is known about their effects on microbial communities. Since many microbes are a good indicator of ecosystem health and respond sensitively to stressors, we expected a shift of the microbial communities under exposure to insecticide and increased fine sediment deposition, at least due to indirect top-down effects caused by the reduced abundance of benthic macroinvertebrates that typically graze on cotton strips. But it is also possible that direct or indirect effects of the stressors on microbes were too low to cause a sufficient shift in microbial communities for taxonomy-based stressor predictions or even that increased fine sediment deposition was beneficial for microbial communities because it provided additional surface habitat for microbes or stimulated organic matter decomposition through physical abrasion of the cotton strips. Therefore, our observed insufficient SPP could also be a consequence of stressor choice rather than limitations of sequencing depth or machine learning, especially since other studies show good performance of machine learning models for environmental assessments based on amplicon sequencing (Cordier et al., 2017, 2018; Gerhard and Gunsch, 2019; Frühe et al., 2020; Dully et al., 2021).

Smith et al. (2015) showed that the performance of prediction models can highly vary based on the predicted environmental variables (including stressor variables). When they attempted to predict 38 geochemical groundwater variables based on 16S sequencing data, the predicted and actual values of 26 variables significantly correlated with each other while those of 12 variables did not. This was further supported by Hermans et al. (2020), who predicted seven soil variables based on 16S sequencing data, and the correlations between predicted and actual values ranged from weak to strong and were further dependent on the land use type of the investigated samples. This raises the need for more exploratory research using different stressors until machine learning can be broadly applied to ecological assessments that involve many stressors.

Nevertheless, the learning curves generated for our best model indicate that more samples likely would have increased SPP. This result is promising because it shows that further sampling likely would have revealed subtle yet distinctive community shifts that would have allowed for better predictions without requiring further knowledge about the direct or indirect effects of the stressors on microbes, which further highlights the potential of machine learning for HTS-based environmental assessments given sufficient sampling size.

We have only investigated the taxonomic information generated by metagenomics and total RNA-Seq, but both methods also generate information on functional diversity (metagenomics) and differential gene expression (total RNA-Seq). This information can also be integrated, which is why omics-based methods are gaining increased attention for environmental assessments (Uyaguari-Diaz et al., 2016; Leese et al., 2018; Cordier et al., 2019, 2021), and it remains to be tested to what extent SPP can be increased by integrating taxonomical and functional information.

4.5 Conclusion

We demonstrate that sequencing and data-processing methods have a substantial impact on environmental stressor prediction when applying machine learning to taxonomically assigned HTS data. Omics-based methods detected much more taxa than amplicon sequencing, and while this is considered an advantage, amplicon sequencing, specifically 16S sequencing, outperformed all other sequencing methods in terms of stressor prediction performance (SPP). However, the best observed SPP for 16S sequencing was only moderate to good, meaning that further improvements are necessary to meet the required standard for applied ecological assessments. Nevertheless, learning curves indicated that more samples would likely have increased SPP, demonstrating the potential for further research. Omics-based methods performed poorly, possibly due to insufficient sequencing depth or a too shallow taxonomic resolution of crucial taxa, but given that other studies demonstrated the potential of omics-based methods in combination with machine learning, further omics-based ecological research is required to show if this approach holds potential for environmental stressor predictions. Data types had no impact on SPP while feature selection significantly improved SPP for omics-based methods but not for amplicon sequencing, and if similar studies confirm these results, then this would strongly favor the application of amplicon sequencing over omics-based methods for environmental assessments. However, we only investigated taxonomic information, but omics-based methods also generate functional information, and it remains to be tested whether the integration of taxonomic and functional information can further improve omics-based environmental assessments.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJNA903104.

Author contributions

DB, LM, MB, and FL designed the experiment. DB, LM, and MB conducted the experiment and collected the samples. DB, LM, and CH processed the samples. DB and CH processed the sequencing data. CH and DT analyzed the data. CH drafted the manuscript. All authors read and approved the final manuscript.

Funding

CH was funded through the Canada First Research Excellence Fund to the program CFREF–Food from Thought at the University of Guelph. DB, MB, and the field experiment were funded through the DFG grants LE 2323/9-1, MA, and SCHA. LM was funded through the Land2Sea project (Aquatic Ecosystem Services in a Changing World, https://land2sea.ucd.ie/; funded under the Joint BiodivERsA-Belmont Forum call and the DFG) and the DFG project LE2323/9-1/MA XXXXX 418091530.

Acknowledgments

We are grateful to Christoph Mayer, Peter Haase, and Ralf Schäfer for their support during the grant application, and we thank Verena Schreiner for performing the pesticide analysis and Romana Salis for performing the taxonomic annotation of amplicon sequencing data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1217750/full#supplementary-material

References

Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R. H., et al. (2021). *UNITE general FASTA release for eukaryotes*.

Ai, D., Pan, H., Han, R., Li, X., Liu, G., and Xia, L. C. (2019). Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. *Genes (Basel)* 10:112. doi: 10.3390/genes10020112

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

Auladell, A., Barberán, A., Logares, R., Garcés, E., Gasol, J. M., and Ferrera, I. (2022). Seasonal niche differentiation among closely related marine bacteria. *ISME J.* 16, 178–189. doi: 10.1038/s41396-021-01053-2

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

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* 41, 1633–1644. doi: 10.1038/s41587-023-01688-w

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

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., and Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS One* 15, 1–21. doi: 10.1371/journal.pone.0228899

Buchner, D., Beermann, A. J., Laini, A., Rolauffs, P., Vitecek, S., Hering, D., et al. (2019). Analysis of 13,312 benthic invertebrate samples from German streams reveals minor deviations in ecological status class between abundance and presence/absence data. *PLoS One* 14, 1–18. doi: 10.1371/journal.pone.0226547

Buchner, D., Beermann, A. J., Leese, F., and Weiss, M. (2021). Cooking small and large portions of "biodiversity-soup": miniaturized DNA metabarcoding PCRs perform as good as large-volume PCRs. *Ecol. Evol.* 11, 9092–9099. doi: 10.1002/ece3.7753

Buchner, D., Macher, T.-H., and Leese, F. (2022). APSCALE: advanced pipeline for simple yet comprehensive analyses of DNA Meta-barcoding data. *Bioinformatics* 7, 1–3. doi: 10.1093/bioinformatics/btac588

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4516–4522. doi: 10.1073/pnas.1000080107

Caruso, V., Song, X., Asquith, M., and Karstens, L. (2019). Performance of microbiome sequence inference methods in environments with varying biomass. *mSystems* 4:e00163. doi: 10.1128/msystems.00163-18

Chan, C. S., Chan, K. G., Tay, Y. L., Chua, Y. H., and Goh, K. M. (2015). Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Front. Microbiol.* 6, 1–15. doi: 10.3389/fmicb.2015.00177

Chang, H. X., Haudenshield, J. S., Bowen, C. R., and Hartman, G. L. (2017). Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Front. Microbiol.* 8, 1–11. doi: 10.3389/fmicb.2017.00519

Chen, T., and Guestrin, C. (2016). *XGBoost: a scalable tree boosting system*. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York. 785–794.

Choe, Y. H., Kim, M., and Lee, Y. K. (2021). Distinct microbial communities in adjacent rock and soil substrates on a high Arctic Polar Desert. *Front. Microbiol.* 11, 1–15. doi: 10.3389/fmicb.2020.607396

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2021). Ecosystems monitoring powered by environmental genomics: a review of current strategies with an implementation roadmap. *Mol. Ecol.* 30, 2937–2958. doi: 10.1111/mec.15472

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J. A., Ouadahi, A., Martins, C., et al. (2017). Predicting the ecological quality status of marine environments from eDNA

Metabarcoding data using supervised machine learning. *Environ. Sci. Technol.* 51, 9118–9126. doi: 10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., and Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* 18, 1381–1391. doi: 10.1111/1755-0998.12926

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

Crisci, C., Ghattas, B., and Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Model.* 240, 113–122. doi: 10.1016/j.ecolmodel.2012.03.001

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Guèze, M., Agard, J., et al. (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*. Bonn, Germany.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14, 671–683. doi: 10.1093/bib/bbs046

Douglas, G. M., Hansen, R., Jones, C. M. A., Dunn, K. A., Comeau, A. M., Bielawski, J. P., et al. (2018). Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn' s disease. *Microbiome* 6, 1–12. doi: 10.1186/s40168-018-0398-3

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., et al. (2021). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture–an inter-laboratory study. *Ecol. Indic.* 121:7049. doi: 10.1016/j.ecolind.2020.107049

Farinella, R., Rizzato, C., Bottai, D., Bedini, A., Gemignani, F., Landi, S., et al. (2022). Maternal anthropometric variables and clinical factors shape neonatal microbiome. *Sci. Rep.* 12, 1–10. doi: 10.1038/s41598-022-06792-6

Frey, B., Rime, T., Phillips, M., Stierli, B., Hajdas, I., Widmer, F., et al. (2016). Microbial diversity in European alpine permafrost and active layers. *FEMS Microbiol. Ecol.* 92, 1–17. doi: 10.1093/femsec/fiw018

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., et al. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* 8:312. doi: 10.1038/s41467-017-01312-x

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., et al. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Mol. Ecol.* 30, 2988–3006. doi: 10.1111/mec.15434

Gerhard, W. A., and Gunsch, C. K. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environ. Int.* 124, 312–319. doi: 10.1016/j.envint.2018.12.038

Ghannam, R. B., and Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* 19, 1092–1107. doi: 10.1016/j.csbj.2021.01.028

Glasl, B., Bourne, D. G., Frade, P. R., Thomas, T., Schaffelke, B., and Webster, N. S. (2019). Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome* 7, 1–13. doi: 10.1186/s40168-019-0705-7

Glassman, S. I., and Martiny, J. B. H. (2018). Broadscale ecological patterns are robust to use of exact. *mSphere* 3:e00148. doi: 10.1128/mSphere.00148-18

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8, 1–6. doi: 10.3389/fmicb.2017.02224

Good, S. P., Urycki, D. R., and Crump, B. C. (2018). Predicting hydrologic function with aquatic gene fragments. *Water Resour. Res.* 54, 2424–2435. doi: 10.1002/2017WR021974

Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55. doi: 10.1038/s41580-021-00407-0

Haase, P., Frenzel, M., Klotz, S., Musche, M., and Stoll, S. (2016). The long-term ecological research (LTER) network: relevance, current status, future perspective and examples from marine, freshwater and terrestrial long-term observation. *Ecol. Indic.* 100, 1–3. doi: 10.1016/j.ecolind.2016.01.040

Hacilar, H., Nalbantoglu, O. U., and Bakir-Gungor, B. (2018). Machine learning analysis of inflammatory bowel disease-associated metagenomics dataset. UBMK 2018-3rd Int. *Conf. Comput. Sci. Eng.* 2018, 434–438. doi: 10.1109/UBMK.2018.8566487

Harris, C. R., Millman, K. J., Van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hempel, C. A., Carson, S. E. E., Elliott, T. A., Adamowicz, S. J., and Steinke, D. (2023). Reconstruction of small subunit ribosomal RNA from high-throughput sequencing data: a comparative study of metagenomics and total RNA sequencing. *Methods Ecol. Evol.* 14, 2049–2064. doi: 10.1111/2041-210X.14149

Hempel, C. A., Wright, N., Harvie, J., Hleap, J. S., Adamowicz, S. J., and Steinke, D. (2022). Metagenomics versus total RNA sequencing: most accurate data-processing tools, microbial identification accuracy, and perspectives for freshwater assessments. *Nucleic Acids Res.* 50, 9279–9293. doi: 10.1093/nar/gkac689

Hermans, S. M., Buckley, H. L., Case, B. S., Curran-Cournane, F., Taylor, M., and Lear, G. (2020). Using soil bacterial communities to predict physico-chemical variables and soil quality. *Microbiome* 8, 1–13. doi: 10.1186/s40168-020-00858-1

Joos, L., Beirinckx, S., Haegeman, A., Debode, J., Vandecasteele, B., Baeyen, S., et al. (2020). Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. *BMC Genomics* 21, 1–17. doi: 10.1186/s12864-020-07126-4

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., et al. (2021). Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam co, Tibet: the effect of sample-size and bioinformatics on the identified communities. *Ecol. Indic.* 121:7070. doi: 10.1016/j.ecolind.2020.107070

Kask, O., Kyman, S., Conn, K. A., Gormley, J., Gardner, J., Johns, R. A., et al. (2020). Environmental exposures influence nasal microbiome composition in a longitudinal study of division I collegiate athletes. *BioRxiv* 2020:946475. doi: 10.1101/2020.02.13.946475

Khachatryan, L., De Leeuw, R. H., Kraakman, M. E. M., Pappas, N., Te Raa, M., Mei, H., et al. (2020). Taxonomic classification and abundance estimation using 16S and WGS–A comparison using controlled reference samples. *Forensic Sci. Int. Genet.* 46:102257. doi: 10.1016/j.fsigen.2020.102257

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422. doi: 10.1038/s41579-018-0029-9

Knowles, S. C. L., Eccles, R. M., and Baltrūnaitė, L. (2019). Species identity dominates over environment in shaping the microbiota of small mammals. *Ecol. Lett.* 22, 826–837. doi: 10.1111/ele.13240

Koh, H., Li, Y., Zhan, X., Chen, J., and Zhao, N. (2019). A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. *Front. Genet.* 10, 1–14. doi: 10.3389/fgene.2019.00458

Kubiszewski, I., Costanza, R., Anderson, S., and Sutton, P. (2017). The future value of ecosystem services: global scenarios and national implications. *Ecosyst. Serv.* 26, 289–301. doi: 10.1016/j.ecoser.2017.05.004

Lanzén, A., Mendibil, I., Borja, A., and Saez, L. A. (2020). A microbial mandala for environmental monitoring – predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Mol. Ecol.* 30, 2969–2987. doi: 10.1111/mec.15489

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., and Carissimi, C. (2018). Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. *Omi. A J. Integr. Biol.* 22, 248–254. doi: 10.1089/omi.2018.0013

Laursen, M. F., Dalgaard, M. D., and Bahl, M. I. (2017). Genomic GC-content affects the accuracy of 16S rRNA gene sequencing bsed microbial profiling due to PCR bias. *Front. Microbiol.* 8, 1–8. doi: 10.3389/fmicb.2017.01934

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

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, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

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.* 7:987. doi: 10.3389/fmicb.2016.00987

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

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

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., et al. (2013). Meta-analyses of studies of the human microbiota. *Genome Res.* 23, 1704–1714. doi: 10.1101/gr.151803.112

Mack, L., Buchner, D., Brasseur, M. V., Leese, F., Piggott, J. J., Tiegs, S. D., et al. (2022). *Fine sediment and the insecticide chlorantraniliprole inhibit organic matter decomposition in streams through different pathways*. Freshw, Biol.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification,

Disease Prediction and Treatment. *Front. Microbiol.* 12:4511. doi: 10.3389/fmicb.2021.634511

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

McLaren, M. R., and Callahan, B. J. (2021). *Silva 138.1 prokaryotic SSU taxonomic training data formatted for DADA2*.

Meisel, J. S., Hannigan, G. D., Tyldsley, A. S., SanMiguel, A. J., Hodkinson, B. P., Zheng, Q., et al. (2016). Skin microbiome surveys are strongly influenced by experimental design. *J. Invest. Dermatol.* 136, 947–956. doi: 10.1016/j.jid.2016.01.016

Mirtl, M., Borer, E. T., Djukic, I., Forsius, M., Haubold, H., Hugo, W., et al. (2018). Genesis, goals and achievements of long-term ecological research at the global scale: a critical review of ILTER and future directions. *Sci. Total Environ.* 626, 1439–1462. doi: 10.1016/j.scitotenv.2017.12.001

Muletz Wolz, C. R., Yarwood, S. A., Campbell Grant, E. H., Fleischer, R. C., and Lips, K. R. (2018). Effects of host species and environment on the skin microbiome of plethodontid salamanders. *J. Anim. Ecol.* 87, 341–353. doi: 10.1111/1365-2656.12726

Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., and Massana, R. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol. Ecol. Resour.* 20, 718–731. doi: 10.1111/1755-0998.13147

Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics* 19, 1–13. doi: 10.1186/s12859-018-2205-3

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

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

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

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19, 1–17. doi: 10.1186/s12864-018-4637-6

Pettorelli, N., Graham, N. A. J., Seddon, N., Da Cunha, M., Bustamante, M., Lowton, M. J., et al. (2021). Time to integrate global climate change and biodiversity science-policy agendas. *J. Appl. Ecol.* 58, 2384–2393. doi: 10.1111/1365-2664.13985

Philippot, L., Andersson, S. G. E., Battin, T. J., Prosser, J. I., Schimel, J. P., Whitman, W. B., et al. (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* 8, 523–529. doi: 10.1038/nrmicro2367

Piggott, J. J., Salis, R. K., Lear, G., Townsend, C. R., and Matthaei, C. D. (2015). Climate warming and agricultural stressors interact to determine stream periphyton community composition. *Glob. Chang. Biol.* 21, 206–222. doi: 10.1111/gcb.12661

Pinto, A. J., and Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* 7:3093. doi: 10.1371/journal.pone.0043093

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

Reback, J., McKinney, W. J., Van Den Bossche, J., Augspurger, T., Cloud, P., et al. (2021). *Pandas-dev/pandas: Pandas 1.3.5*.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *Peer J* 2016, 1–22. doi: 10.7717/peerj.2584

Roy, J., Mazel, F., Sosa-Hernández, M. A., Dueñas, J. F., Hempel, S., Zinger, L., et al. (2019). The relative importance of ecological drivers of arbuscular mycorrhizal fungal distribution varies with taxon phylogenetic resolution. *New Phytol.* 224, 936–948. doi: 10.1111/nph.16080

Sagova-Mareckova, M., Boenigk, J., Bouchez, A., Cermakova, K., Chonova, T., Cordier, T., et al. (2021). Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Res.* 191:116767. doi: 10.1016/j.watres.2020.116767

Shah, N., Tang, H., Doak, T. G., and Ye, Y. (2010). Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac. Symp. Biocomput.* 2011, 165–176. doi: 10.1142/9789814335058_0018

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

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–e00315. 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

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., and Vasselon, V. (2019). Diatom DNA Metabarcoding for biomonitoring: strategies to avoid major taxonomical and Bioinformatical biases limiting molecular indices capacities. *Front. Ecol. Evol.* 7, 1–15. doi: 10.3389/fevo.2019.00409

Tavalire, H. F., Christie, D. M., Leve, L. D., Ting, N., Cresko, W. A., and Bohannan, B. J. M. (2021). Shared environment and genetics shape the gut microbiome after infant adoption. *MBio* 12:548. doi: 10.1128/mBio.00548-21

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., et al. (2022). Best practices in metabarcoding of fungi: from experimental design to results. *Mol. Ecol.* 31, 2769–2795. doi: 10.1111/mec.16460

Tessler, M., Neumann, J. S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L. F. M., et al. (2017). Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.* 7, 1–14. doi: 10.1038/s41598-017-06665-3

The Scikit-Bio Development Team. (2020). *Scikit-bio: A bioinformatics library for data scientists, students, and developers*. Available at: http://scikit-bio.org.

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:e2527. doi: 10.1371/journal.pone.0002527

Uyaguari-Diaz, M. I., Chan, M., Chaban, B. L., Croxen, M. A., Finke, J. F., Hill, J. E., et al. (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* 4, 1–19. doi: 10.1186/s40168-016-0166-1

Van Rossum, G., and Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

Vera-Gargallo, B., Chowdhury, T. R., Brown, J., Fansler, S. J., Durán-Viseras, A., Sánchez-Porro, C., et al. (2019). Spatial distribution of prokaryotic communities in hypersaline soils. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-018-38339-z

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., and Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* 3, 1–11. doi: 10.1186/s40168-015-0087-4

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Witten, I. H., and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. *2nd Edn*. San Francisco: Elsevier Inc.

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:e1000667. doi: 10.1371/journal.pcbi.1000667

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., et al. (2018). Metagenomics biomarkers selected for prediction of three different diseases in Chinese population. *Biomed. Res. Int.* 2018:36257. doi: 10.1155/2018/2936257

WWF. (2020). *Living planet report 2020-bending the curve of biodiversity loss*. Gland, Switzerland: WWF.

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

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., Pruesse, E., Quast, C., and Glöckner, F. O. (2011). Analysis of 23S rRNA genes in metagenomes - a case study from the Global Ocean sampling expedition. *Syst. Appl. Microbiol.* 34, 462–469. doi: 10.1016/j.syapm.2011.04.005

Zizka, V. M. A., Elbrecht, V., Macher, J. N., and Leese, F. (2019). Assessing the influence of sample tagging and library preparation on DNA metabarcoding. *Mol. Ecol. Resour.* 19, 893–899. doi: 10.1111/1755-0998.13018