scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Metatranscriptomes of activated sludge microbiomes from saline wastewater treatment plant

Asala Mahajna[1,2] ✉, Bert Geurkink[1], Ranko Gacesa[3,4], Karel J. Keesman[5], Gert-Jan W. Euverink[2] & Bayu Jayawardhana[2]

The activated sludge microbiome (ASM) drives the biological wastewater treatment process in wastewater treatment plants. It has been established in the literature that the ASM is characterized by a high degree of taxonomic and metabolic diversity. However, meta-omics datasets have been derived from domestic wastewater treatment plants with little attention to saline wastewater treatment plants (SWWTP). Existing knowledge of how activated sludge microorganisms impact water quality, interrelate within habitat networks, and respond to environmental perturbations remains limited. Here we present datasets of the metatranscriptomes of SWWTP in The Netherlands, coupled with process data. The dataset represents a two-year and four-month time series of data collected from 2014 to 2017, with samples taken at approximately monthly intervals from the facultative zone in the activated sludge process of an SWWTP. In total, 32 activated sludge samples were analyzed. This dataset can be used to enhance understanding of the unique microbiome composition in SWWTPs, its dynamic responses to environmental variables, and the metabolic functions within the ASM.

Background & Summary

The activated sludge microbiome (ASM) is the driving force behind the biological wastewater treatment processes in wastewater treatment plants (WWTPs). It is well established that a wealth of taxonomic and metabolic diversity is present in the ASM[1–3]. However, a comprehensive understanding of how these microorganisms exert influence on water quality, establish intricate relationships within their habitat network, and respond to environmental perturbations is yet to be fully elucidated. This knowledge gap underscores the need for research in this domain. The advent of next-generation sequencing technologies (NGS) in the water sector offers a unique opportunity to gain a more profound insight into the water microbiome[4].

The integration of NGS data into the study of activated sludge has significantly advanced our understanding of microbial dynamics within wastewater treatment systems. In particular, this technology has enabled the precise identification of bacterial species and their degradation capabilities, shedding light on the complex processes involved in breaking down pollutants and organic matter[5,6]. By providing a comprehensive view of the metabolic pathways active within these communities, NGS has revealed the multifaceted roles of microorganisms in nutrient cycling and organic matter decomposition[7–10]. These discoveries have not only enhanced our fundamental understanding of microbial ecology but have also inspired the development of innovative resource recovery strategies, further promoting a paradigm shift in how wastewater treatment plants can function as bio-refineries[11,12]. Moreover, NGS has been pivotal in understanding how microbial communities in activated sludge respond to environmental perturbations, providing critical insights necessary for maintaining the stability and efficiency of wastewater treatment processes under variable conditions[13–16]. The technology has also illuminated the mechanisms underlying the spread of antibiotic resistance within these systems, highlighting

[1]Wetsus – European Centre of Excellence for Sustainable Water Technology, Oostergoweg 9, 8911 MA, Leeuwarden, The Netherlands. [2]Engineering and Technology Institute Groningen, Faculty of Science and Engineering, University of Groningen, Nijenborgh 4, 9747 AG, Groningen, The Netherlands. [3]Department of Genetics, University of Groningen and University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV, Groningen, The Netherlands. [4]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV, Groningen, The Netherlands. [5]Mathematical and Statistical Methods – Biometris, Wageningen University, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands. ✉e-mail: a.mahajna@rug.nl

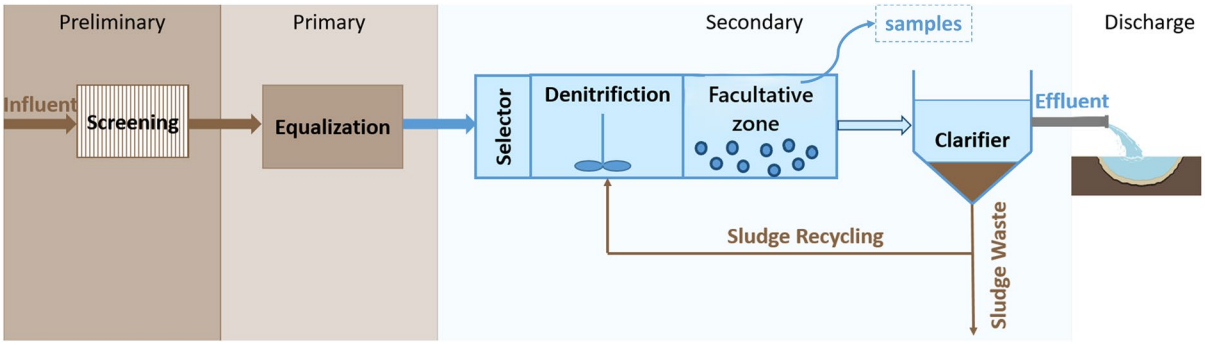

Fig. 1 Schematic representation of the treatment process at Oosterhorn SWWTP showing a preliminary treatment step based on screens, a primary treatment step based on an equalization tank, and a secondary treatment based on an activated sludge tank and a settling tank.

the intricate interactions between resistance genes and mobile genetic elements[17–20]. This information is crucial for developing effective measures to mitigate the dissemination of antibiotic resistance from treatment plants to natural environments, addressing a major public health concern. Potential measures may include optimization of microbial growth conditions and bio-augmentation strategies, resulting in enhanced treatment efficiency and environmental protection[21,22].

This paper presents a dataset from a sampling campaign aimed at advancing our understanding of microbial gene expression and metabolic pathways in activated sludge, which play a crucial role in processes such as pollutant degradation and nutrient cycling. While extensive studies on the activated sludge microbiome using 16S rRNA sequencing have contributed to the development of MIDAS (Microbial Database for Activated Sludge)[23], a reference database specific to this ecosystem, 16S rRNA sequencing alone is insufficient for understanding metabolic functions. This technique focuses on ribosomal RNA genes for taxonomic classification, providing limited insights into the functional potential of microbial communities. Although metagenomics offers a broader view of microbial metabolic pathways by sequencing all genetic material, it only reveals the genetic potential without indicating active metabolic processes. Furthermore, 16S rRNA sequencing cannot capture real-time gene expression or metabolic activity, nor can it detect low-abundance functional genes or pathways. To gain a more comprehensive understanding of microbial metabolism, particularly in complex systems, metatranscriptomics, which assesses gene expression, is essential. Thus, this study provides metatranscriptome data from saline-activated sludge to offer a more accurate representation of its functional capabilities.

Furthermore, the campaign enables the exploration of the differences in microbial community structure and function between SWWTPs and conventional domestic systems. In addition to microbiome data, operational process data from the SWWTP were gathered to assess the impact of operational parameters on the structure and functionality of the ASM.

In this context, we present metatranscriptome datasets derived from a SWWTP located in The Netherlands. The sampling site is North Water's SWWTP, located at the Oosterhorn industrial park within the harbor district close to the city of Delfzijl in the Netherlands. North Water, a collaborative venture between Evides Industrial Water and Waterbedrijf Groningen utility, constructed the SWWTP and maintains it under a Design, Build, Finance & Operate (DBFO) arrangement. The industrial wastewater arriving at the SWWTP has a pollution equivalent of 35,000 PE (Population Equivalent, a parameter for characterizing industrial wastewater) with a daily peak of 60,000 PE. The hydraulic capacity of the SWWTP is 250 $\frac{m^3}{hour}$. The SWWTP comprises a preliminary treatment step based on screens, a primary treatment step based on an equalization tank, and a secondary treatment based on an activated sludge tank and a settling tank. The activated sludge tank consists of three zones: selector, denitrification zone, and facultative zone. The aeration volume in the facultative zone can vary between 0 and 3,500 $\frac{N_{m^3}}{h}$. The equalization tank has a capacity of 2,500 m^3, while the sludge tank has a capacity of 6,700 m^3, with a sludge age of 3–4 weeks. After treatment, the effluent is discharged into the Eems River through the Zeehavenkanaal, ultimately reaching the Wadden Sea[24,25]. The process scheme is shown in Fig. 1.

The sampling campaign was conducted at the Oosterhorn SWWTP from 2014 until 2017. The aim of the sampling campaign was to gain a deeper insight into the activated sludge microbial community composition and function in a saline environment. In total, 32 samples were collected from the facultative zone in the activated sludge process in the SWWTP where 9% of the samples were collected in 2014 (n = 3), 44% of the samples were collected in 2015 (n = 14), 44% of the samples were collected in 2016 (n = 14), and 3% of the samples were collected in 2017 (n = 1). Samples were collected 2–4 weeks apart. The exact date of collection of each sample is outlined in Table 1.

The genetic datasets are augmented with concurrent physical, chemical, and biological measurements obtained from diverse locations within the SWWTP[26]. Similar physical, chemical, and biological *ex-situ* laboratory analyses were conducted on wastewater samples from the influent and effluent streams for the SWWTP. Table 2 summarizes the types of measures collected at the influent and effluent of the SWWTP.

Additionally, Table 3 summarizes data collected hourly from the activated sludge process in the SWWTP.

Additional process measures were also taken on a weekly basis, as detailed in Table 4.

Date	Code	Sample Description	Year	Week
14-Oct-14	4717_003	ZAWZI_Oosterhorn_AT_20141014	2014	42
14-Nov-14	4717_008	ZAWZI_Oosterhorn_AT_20141114	2014	46
17-Dec-14	4717_010	ZAWZI_Oosterhorn_AT_20141217	2014	51
07-Jan-15	4717_012	ZAWZI_Oosterhorn_AT_20150107	2015	2
22-Jan-15	4717_013	ZAWZI_Oosterhorn_AT_20150122	2015	4
04-Mar-15	4717_014	ZAWZI_Oosterhorn_AT_20150304	2015	10
02-Apr-15	4717_016	ZAWZI_Oosterhorn_AT_20150402	2015	14
05-May-15	4717_018	ZAWZI_Oosterhorn_AT_20150505	2015	19
28-May-15	4717_020	ZAWZI_Oosterhorn_AT_20150528	2015	22
23-Jun-15	4717_022	ZAWZI_Oosterhorn_AT_20150623	2015	26
21-Jul-15	4717_024	ZAWZI_Oosterhorn_AT_20150721	2015	30
19-Aug-15	4717_026	ZAWZI_Oosterhorn_AT_20150819	2015	34
18-Sep-15	4717_028	ZAWZI_Oosterhorn_AT_20150918	2015	38
13-Oct-15	4717_030	ZAWZI_Oosterhorn_AT_20151013	2015	42
11-Nov-15	4717_032	ZAWZI_Oosterhorn_AT_20151111	2015	46
25-Nov-15	4717_033	ZAWZI_Oosterhorn_AT_20151125	2015	48
23-Dec-15	4717_035	ZAWZI_Oosterhorn_AT_20151223	2015	52
14-Jan-16	4717_036	ZAWZI_Oosterhorn_AT_20160114	2016	3
03-Feb-16	4717_037	ZAWZI_Oosterhorn_AT_20160203	2016	6
09-Mar-16	4717_038	ZAWZI_Oosterhorn_AT_20160309	2016	11
06-Apr-16	4717_040	ZAWZI_Oosterhorn_AT_20160406	2016	15
20-Apr-16	4717_041	ZAWZI_Oosterhorn_AT_20160420	2016	17
18-May-16	4717_042	ZAWZI_Oosterhorn_AT_20160518	2016	21
01-Jun-16	4717_043	ZAWZI_Oosterhorn_AT_20160601	2016	23
29-Jun-16	4717_044	ZAWZI_Oosterhorn_AT_20160629	2016	27
20-Jul-16	4717_047	ZAWZI_Oosterhorn_AT_20160720	2016	30
18-Aug-16	4717_051	ZAWZI_Oosterhorn_AT_20160818	2016	34
12-Sep-16	4717_054	ZAWZI_Oosterhorn_AT_20160912	2016	38
12-Oct-16	4717_058	ZAWZI_Oosterhorn_AT_20161012	2016	42
17-Nov-16	4717_081	ZAWZI_Oosterhorn_AT_20161117	2016	47
22-Dec-16	4717_082	ZAWZI_Oosterhorn_AT_20161222	2016	52
09-Feb-17	4717_083	ZAWZI_Oosterhorn_AT_20170209	2017	6

Table 1. Sampling time and sample description of the activated sludge process' sampling campaign at the Oosterhorn SWWTP.

In addition, glycerol and methanol content in the influent were measured daily during the sampling period. Also, measurements of the Sludge Volume Index (SVI) after 5, 10, 15, 20, 25 and 30 minutes during the sampling period are also included.

Methods

Step 1: Sample collection and stabilization. Activated sludge samples, each measuring 8 ml in volume, were collected and transferred into 18 ml scintillation vials. Within each vial, a solution comprising 9.5 ml of 96% ethanol and 0.5 ml of 0.1 M sodium citrate at pH 4.2 was added. This composition was designed to stabilize the samples and maintain the integrity of the RNA. Subsequently, the vials were stored at −20 °C to ensure preservation until further analysis.

Step 2: Sample Pre-treatment. A 1 ml volume from each sample was transferred to empty Lysing Matrix E vials for subsequent analysis. Each Lysing Matrix E vial (MPBio-medicals, Solon, OH, USA) containing a portion of the collected samples was centrifuged for 10 minutes at 10,000 × g using an Eppendorf 5424 R centrifuge (Eppendorf AG, Hamburg, Germany). This centrifugation step was performed to concentrate the cells before proceeding with the RNA extraction process in Step 3.

Step 3: RNA extraction. RNA extraction was carried out using a modified protocol based on the FastRNA ProSoil Direct kit (MPBio-medicals, Solon, OH, USA). The extraction procedure involved several steps. First, 700 µl of RNApro Soil Lysis Solution was added to a Lysis Matrix E vial containing the sample pellet. Next, 350 µl of acidified phenol/chloroform from the kit was introduced, and the mixture was homogenized for 45 seconds at a setting of 6.0 using the MP Biomedicals FastPrep-24™ 5 G bead-beating grinder. This process ensured thorough disruption of the sample and release of RNA from the cellular matrix.

Following homogenization, 600 µl of acidified phenol/chloroform was added to the lysate, which was then centrifuged at 10,000 × g for 3 minutes at 4 °C to separate cellular debris. The upper aqueous phase, which

Acronyms	Measure	Units
BOD_5	Biological oxygen demand after 5 days	*mg/L*
Cl_2	Chlorine	*mg/L*
COD	Chemical oxygen demand	*mg_O_2/L*
EC	Electrical conductivity	*mS/m*
K^+	Potassium	*mg/L*
TN	Total nitrogen	*mg_N/L*
Na^+	Sodium	*mg/L*
NH_4^+	Ammonium	*mg_N/L*
N_{kj}	Total Kjeldahl nitrogen (TKN)	*mg_N/L*
NO_2^-	Nitrite	*mg_N/L*
NO_3^-	Nitrate	*mg_N/L*
pH	pH	—
PO_4^{3-}	Phosphate in terms of phosphorus (P) content	*mg_P/L*
PO_4^{3-}o	Orthophosphate in terms of phosphorus (P) content	*mg_P/L*
PO_4^{3-}o	Orthophosphate as a whole compound	*mg/L*
SO_4^{2-}	Sulfate	*µg/L*
TOD	Total oxygen demand	*mg_O_2/L*
TSS	Total suspended solids	*mg/L*

Table 2. Types of water quality analysis performed on the influent and effluent samples.

Acronyms	Measure	Units
Return_sludge	Flow rate of the return sludge stream	m^3/h
Vol_aeration	Volume of aeration	Nm^3/h
Inflow	Inflow	m^3/h
Blowers_capacity	Capacity of blowers	%
Eff_EC	Effluent electrical conductivity	*mS/cm*
Eff_pH	Effluent pH	—
Eff_Turbidity	Effluent turbidity	—
DW_AT	Dry weight of the activated sludge in the tank also known as mixed liquor suspended solids (MLSS)	*g/L*
pH_denit	pH in the denitrification zone	—
O2_facultative	Dissolved oxygen in the facultative process	*mg/L*
O2_nitrification_zone_2	Dissolved oxygen in nitrification zone 1 in the facultative process	*mg/L*
O2_nitrification_zone_1	Dissolved oxygen in nitrification zone 2 in the facultative process	*mg/L*

Table 3. The hourly process data collected from Oosterhorn SWWTP during the sampling period.

Acronyms	Measure	Units
Temperature	Average temperature	°C
Total_P	Total phosphorus in the inflow stream	*kg/day*
Sludge_load	Sludge load	*kg_COD/kg_DW/day*
Sludge_production	Sludge production	*ton/week*
Yield_Sludge	Sludge yield	*kg_DW/kg_COD_removed*

Table 4. Additional weekly process data of Oosterhorn SWWTP during the sampling period.

contains the RNA, was carefully transferred to a new 1.5 ml collection tube. The cleared lysate was then transferred to a spin column designed for selective nucleic acid binding. The column was washed to remove impurities such as proteins and salts, and RNA was eluted using nuclease-free water from the kit. This elution step was performed twice, with each elution yielding 50 µl of purified RNA.

The resulting RNA eluate underwent treatment with DNase (Promega, RQ1 RNase-Free DNase, Catalog number: M6101) to minimize the co-purification of DNA within the RNA sample. This enzymatic treatment effectively degrades any contaminating DNA molecules, thereby reducing the risk of false positive signals during sequencing.

Subsequently, a column-based purification step was conducted using the Zymo RNA Clean & Concentrator-5 kit (Catalog number: R1015). This purification process further enhances the quality of the RNA sample by removing impurities and residual reagents.

Following purification, the RNA was eluted in a total volume of 30 µl nuclease-free water to concentrate and facilitate accurate quantification. The RNA in the resulting sample was quantified using the Qubit™ RNA High

Sensitivity (HS), Broad Range (BR), and Extended Range (XR) Assay Kit (Catalog number: Q32852). This assay provides precise and reliable quantification of RNA across a broad range of concentrations, ensuring an accurate assessment of RNA yield.

Step 4: cDNA synthesis. After obtaining the RNA eluate, 13 µl of it was utilized to synthesize the first complementary DNA (cDNA) strand through a reverse transcriptase step using the Bioline SensiFAST cDNA Synthesis kit (Catalog number: BIO-65053). Random primers (hexamers) served as primers for this step. The synthesis program included an initial incubation of 5 minutes at 25 °C, followed by 30 minutes at 42 °C for reverse transcription. The reaction was terminated by incubating at 85 °C for 5 minutes to inactivate the reverse transcriptase, followed by cooling to 4 °C.

Following the first strand synthesis, the second cDNA strand was synthesized using Klenow DNA polymerase (Promega, DNA Polymerase I Large (Klenow) Fragment, Catalog number: M2201). Random primers (hexamers) were again used as primers. The synthesis program included an initial incubation of 5 minutes at 25 °C, followed by 30 minutes at 37 °C for the Klenow DNA polymerase reaction. The reaction was terminated by incubating at 75 °C for 5 minutes to inactivate the Klenow DNA polymerase while preserving the synthesized double-stranded cDNA. Subsequently, the cDNA was purified using magnetic beads (Beckman Coulter, AMPure XP 60 ML), and eluted in 30 µl of nuclease-free water.

The amount of synthesized cDNA was quantified using the Qubit Fluorometer (Thermo Fisher Scientific) with the Quant-iT™ DNA Assay Kit HS (Q32851). A 20 µl dilution at a concentration of 0.2 ng/µl was prepared as input for the subsequent next-generation sequencing (NGS) library preparation.

Step 5: Library preparation. For library preparation, the samples were processed using the Nextera XT DNA sample preparation kit 24 samples (Illumina, Catalog number: FC-131-1024), along with the default library linkers and adaptors provided by the Nextera XT Index Kit (Illumina, Catalog number: FC-131-1001).

During the initial step of the procedure, the cDNA underwent tagmentation (tagging and fragmentation) facilitated by the Nextera XT transposome. This transposome simultaneously fragmented the input cDNA and added adapter sequences to the ends, enabling subsequent PCR amplification with different indexes added per sample.

For the index-PCR, the following program was employed: 3 minutes at 72 °C, followed by 30 seconds at 95 °C, and then 12 cycles of denaturation at 95 °C for 10 seconds, annealing at 55 °C for 30 seconds, and extension at 72 °C for 30 seconds. Finally, the reaction was concluded with 5 minutes of extension at 72 °C, followed by a hold at 10 °C.

Subsequent to the index-PCR, the cDNA was purified using magnetic beads (Beckman Coulter, AMPure XP 60 ML) according to the protocol provided. The quantity of synthesized cDNA was measured using the Qubit Fluorometer (Thermo Fisher Scientific) with the Quant-iT™ DNA Assay Kit HS (Catalog number: Q32851).

Step 6: Sequencing. The prepared libraries underwent normalization to achieve a suitable concentration for loading onto a flow-cell (Illumina, MiSeq Reagent Kit v2 (300-cycles), Catalog number: MS-102-2002). For each flow cell, the aim was to add a total of 1.1^{10} molecules of DNA. To calculate the amount of DNA required for normalization, an assumption was made regarding the average sequence length, estimated to be 350 base pairs.

Different sequencing runs were processed, and the normalization process ensured that each library contained the appropriate number of cDNA molecules for optimal sequencing performance. Subsequently, the libraries were sequenced using the MiSeq System (Illumina, Catalog number: SY-410-1003).

Step 7: Community profiling. Following quality assessment using FastQC v0.12.1[27] (parameters: default), and MultiQC[28] v1.22.2 (parameters: default), the raw sequence reads underwent pre-processing and community profiling using Galaxy[29]. This included the removal of adapter sequences and trimming low-quality ends using Trimmomatic[30] v0.39. Adapter contamination was removed (parameters: default), and low-quality bases were trimmed using a sliding window approach with a quality threshold of 30, averaging over 4 bases. Reads with insufficient overall quality were discarded. Taxonomic classification of the sequencing reads was performed with Kraken2[31] v.2.1.3 (parameters: minimum base quality = 30, minimum hit groups = 2, confidence score threshold = 0.1)[32] and referencing the RefSeq[33]. PlusPF database (downloaded 15-07-2024). Bayesian probabilities-based re-estimation of abundance was done using Bracken[34] v3.0. Total RNA-Seq data was employed for community profiling to capture a comprehensive view of the microbial community, ensuring the inclusion of all potential taxa, including those with high ribosomal RNA (rRNA) content. This approach was chosen to avoid missing taxa that may be important due to their rRNA abundance. The workflow of the bioinformatics pipeline designed and implemented to generate the abundance data of the ribosomally active community is outlined in Fig. 3. Figure 2 shows the relative abundance of bacterial phyla across the 32 samples.

Step 8: Functional profiling. Following the same quality assessment using FastQC v0.12.1[27] (parameters: default), and MultiQC[28] v1.22.2 (parameters: default), the raw sequence reads underwent pre-processing and analysis using SAMSA2[35] (Simple Analysis of Metatranscriptomes through Sequence Annotation, version 2). The workflow of the bioinformatics pipeline implemented in SAMSA2 is also integrated in Fig. 3. Firstly, SAMSA2 merges paired-end reads from high-throughput sequencing data using PEAR[36] (Paired-End reAd mergeR) v.0.9.8, it removes adapter sequences and trims low-quality ends using Trimmomatic[30] v0.36, and it removes rRNA from the total RNA-seq data for functional analysis using SortMeRNA[37] v.2.1. SAMSA2 performs taxonomic profiling of the metabolically active microbial community and provides functional profiling by annotating genes and

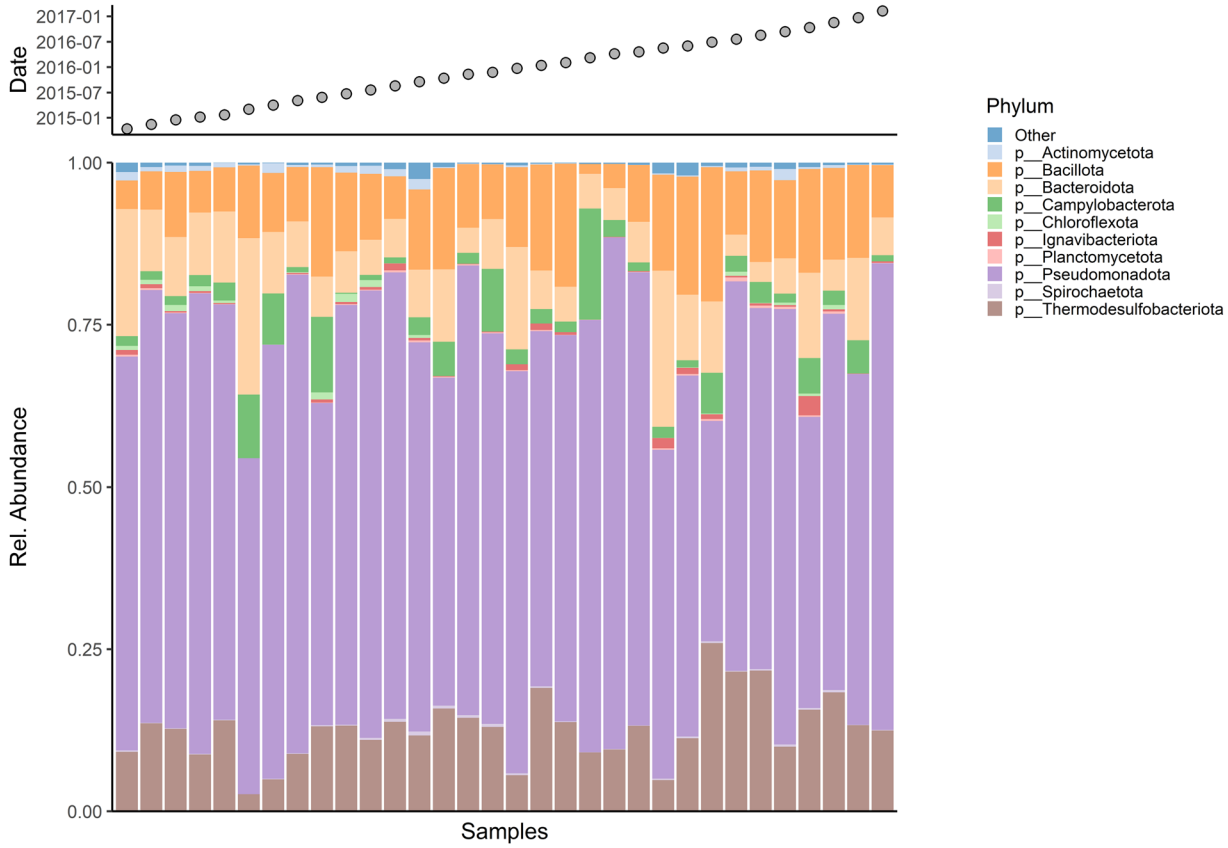

Fig. 2 Relative abundance of the ribosomally active bacterial phyla across the 32 samples.

mapping them to metabolic pathways and functional categories. This is achieved by aligning non-rRNA-seq reads to SEED[38] reference databases (downloaded 18/08/2024) using DIAMOND[39] v.0.8.38.

The data obtained from SAMSA2 yields quantitative information on the abundance of functional genes within the 32 samples of the saline-activated sludge microbiome. SAMSA2 performs functional annotation of enzyme-encoding genes, categorizing them according to the SEED Subsystems framework. SEED Subsystems is a curated database that organizes functional roles into a hierarchical system, facilitating the classification and interpretation of metabolic functions. This hierarchical structure resembles a fractal tree, where the root represents general functional categories, branching out into increasingly specific functions and reaching the enzyme level. The leaves of this fractal tree correspond to highly specific functions and roles of enzymes and proteins[35].

Figure 4 illustrates a heatmap that visualizes the distribution of metabolic functions within the saline-activated sludge microbiome, classified according to the SEED subsystems. Analysis of 32 samples revealed the microbiome's involvement in 31 distinct functional categories, highlighting its considerable functional diversity. To account for compositional biases and mitigate the closure effects inherent in abundance data—where the constant-sum constraint distorts relative abundances and complicates comparisons—a centered log-ratio (CLR) transformation was applied across the samples.

Among the 31 functional categories, a significant proportion of genes were annotated under the "No Hierarchy" category, displaying considerable skewness due to the overrepresentation of certain functional genes. This category, often referred to as functional dark matter, encompasses genes whose functions remain poorly understood, likely due to incomplete or limited databases lacking experimental validation or extensive annotation[40]. The overrepresentation of these genes can obscure the true functional landscape of the microbiome, complicating accurate assessments of its functional potential. To reduce potential bias in data interpretation, a standardization step was applied following the CLR transformation. This step rescales each taxon to have a mean of zero and a variance of one, effectively mitigating the influence of taxa with high abundance. This normalization process enhances the comparability between functional categories of varying abundance, ensuring more balanced and robust analyses of the microbial community's functional profiles.

The resulting heatmap offers valuable insights into the functional diversity of the saline-activated sludge microbiome, highlighting the intricate metabolic roles the microbial community fulfills in wastewater treatment. The data reveals substantial functional heterogeneity, indicative of a highly adaptable microbial community that can dynamically respond to fluctuating environmental conditions while ensuring efficient contaminant removal throughout the year. Despite seasonal variations influencing microbial abundance, composition, and metabolic activity, the microbiome exhibits resilience, demonstrating its capacity to adapt to diverse stresses and consistently uphold its critical role in wastewater treatment processes.

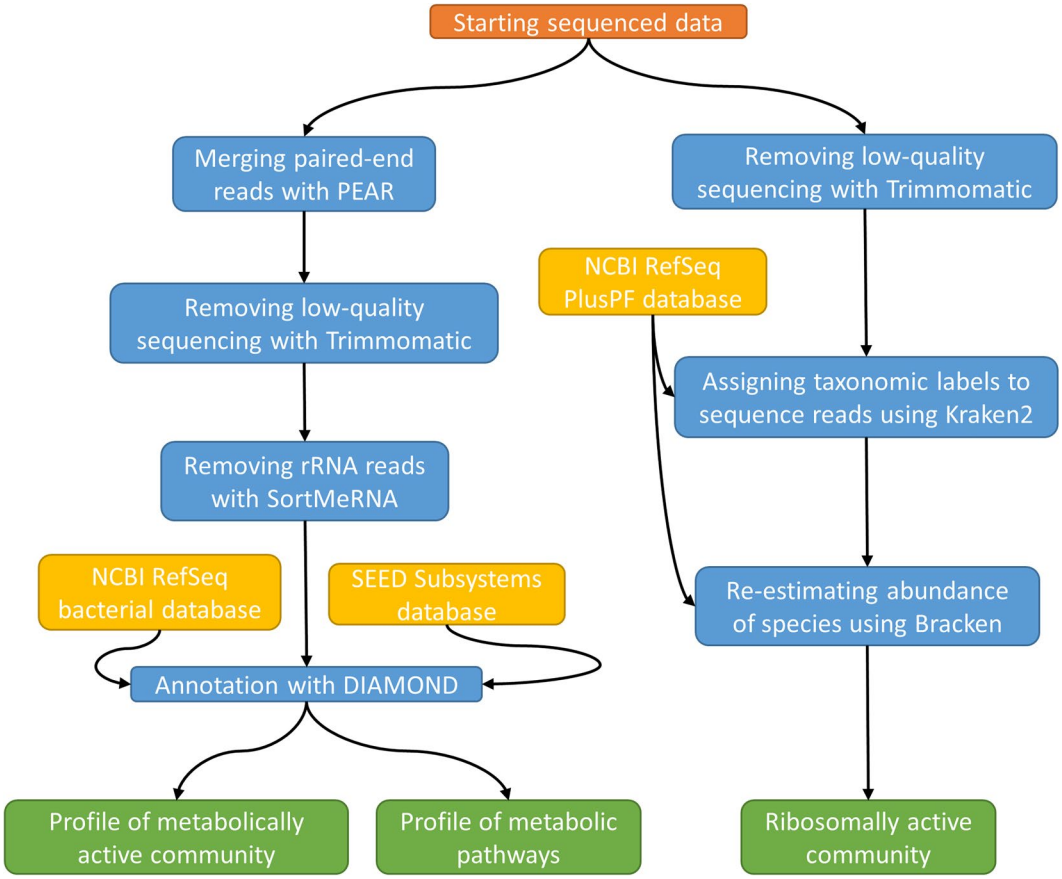

Fig. 3 The combined bio-informatics pipeline.

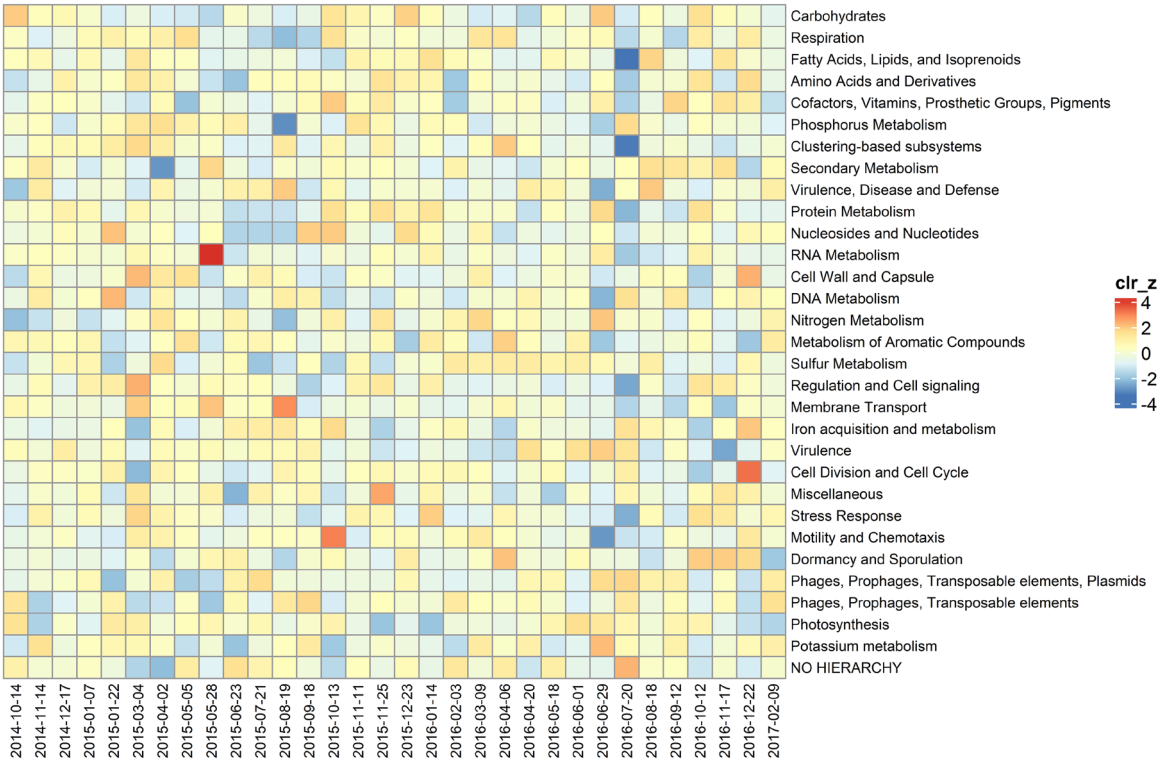

Fig. 4 Heatmap depicting the functional composition across samples, where samples were first subjected to centered log-ratio (CLR) transformation, followed by standardization of functional categories.

In this study, we used two bioinformatics approaches to profile the ASM in the facultative zone of a SWWTP and assess its metabolic functionality. The first approach analyzed ribosomally active communities using the RefSeq[33] PlusPF database, a comprehensive resource for annotating metatranscriptomic data from saline-activated sludge. While not tailored specifically to saline ecosystems or activated sludge microbiomes, the broad coverage of microbial genomes in RefSeq supports accurate taxonomic classification and functional annotation. This general database is particularly useful for complex microbiomes like those in saline wastewater treatment systems, where many taxa may be underrepresented in specialized databases. Using it allows for the inclusion of diverse microbial functions, particularly for uncultured or less well-characterized organisms, ensuring a more comprehensive understanding of the active microbial processes in the system. The second approach focused on metabolically active bacterial communities, identifying microorganisms involved in essential wastewater treatment functions like substrate breakdown, nutrient cycling, and pollutant degradation. This method provides insights into the functional genes and metabolic pathways that drive system efficiency, improving our understanding of the bacterial community's role in wastewater treatment.

Together, these methods provide a comprehensive understanding of microbial dynamics in the wastewater treatment system. Ribosomal activity offers a broad perspective on overall microbial presence, while metabolic activity provides specific insights into functional contributions and biochemical processes. The integration of both approaches allows for a nuanced characterization of the microbiome's role in the wastewater treatment process, highlighting both general microbial activity and specific functional contributions.

Data Records

The 64 raw, unprocessed Illumina sequencing reads (fastq files) for all metatranscriptomes from the pair-end sequencing of the 32 samples have been deposited in the National Center for Biotechnology Information (NCBI) database under the Bioproject ID PRJNA1122484[41] (http://identifiers.org/ncbi/BioProject:PRJNA1122484) and Sequence Read Archive (SRA) project accession number SRP513109[42] (http://identifiers.org/ncbi/insdc.sra:SRP513109). The process data was uploaded to figshare[26] (https://doi.org/10.6084/m9.figshare.27073612.v2).

Technical Validation

Quality assurance of nucleic acids and sequencing libraries. Multiple technical validation steps were conducted to ensure quality control. To start with, quality assurance was conducted on the physical samples as elaborated in the methods section. This was done after RNA extraction using Qubit™ RNA High Sensitivity (HS), Broad Range (BR), and Extended Range (XR) Assay Kit (Catalog number: Q32852). Additionally, a quality assurance step was done to quantify the synthesized cDNA after the library preparation step and just before sequencing, using the Qubit Fluorometer (Thermo Fisher Scientific) with the Quant-iT™ DNA Assay Kit HS (Catalog number: Q32851).

In-Silico validation of sequencing reads. After sequencing, per-sequence quality scores were examined to ensure the quality of the sequencing of each and every sample. Figure 5 shows the per-sequence quality scores. The results show that ~98.67% and ~88.42% of the bases have quality scores of $\geq$20 and $\geq$30, respectively, indicating that sequencing was performed successfully. In line with the output of the Illumina sequencing technology, the forward reads (R1) exhibited higher quality, with ~99.32% of the bases having quality scores of $\geq$20, compared to the reverse reads (R2), with ~98.03% of the bases having quality scores of $\geq$20.

Additionally, a comparative quality assessment of the raw reads of the 32 metatranscriptomes was carried out by combining the general statistics to ensure consistency. This was done using FastQC v0.12.1[27] (parameters: default), and results were aggregated using MultiQC[28] v1.22.2 (parameters: default). Figure 6 summarizes the general statistics of all the sequence data. Overall, the violin plots of the comprehensive statistics of the sequence data present a somewhat symmetrical and smooth profile. The consistent width observed along the length of the plot signifies uniform variability throughout the dataset. Notably, the central mass of the violin plot is clearly delineated, exhibiting a slight skew towards elevated values, thereby implying a positive skewness within the distribution. Furthermore, while isolated outliers are discernible beyond the primary body of the violin plot, their influence on the overarching interpretation is deemed negligible.

Sequencing effort validation. Figure 7 presents the rarefaction curves for the ribosomally active community, with observed species richness plotted as a function of sequencing depth across 32 samples. The rarefaction curves evaluate the sufficiency of sampling for each sample. The consistently plateauing curves indicate that the sequencing effort for all 32 samples was adequate to capture the taxonomic diversity within the saline activated sludge microbiome.

Figure 8 illustrates the rarefaction curves for functional coverage, with the cumulative count of metabolic pathways plotted as a function of sequencing depth across 32 samples. These curves assess the adequacy of sequencing depth for each sample. The consistently plateauing curves suggest that the sequencing efforts were sufficient to capture the diversity of the metabolic pathways present in each sample, indicating that further sequencing would yield diminishing returns in terms of new discoveries. However, variations in the shapes of the curves may indicate differences in sample complexity. To mitigate the impact of sequencing depth variability and improve the reliability of downstream analyses, it is recommended to apply normalization and transformation techniques, such as proportional normalization and compositional data transformations. These steps are particularly important when performing comparative analyses across samples, as demonstrated in Fig. 4.

Ecological validation. Lastly, to ensure the adequacy of the sampling efforts, we evaluated the feature accumulation curves presented in Fig. 9. These curves were constructed for various categories, including metabolic

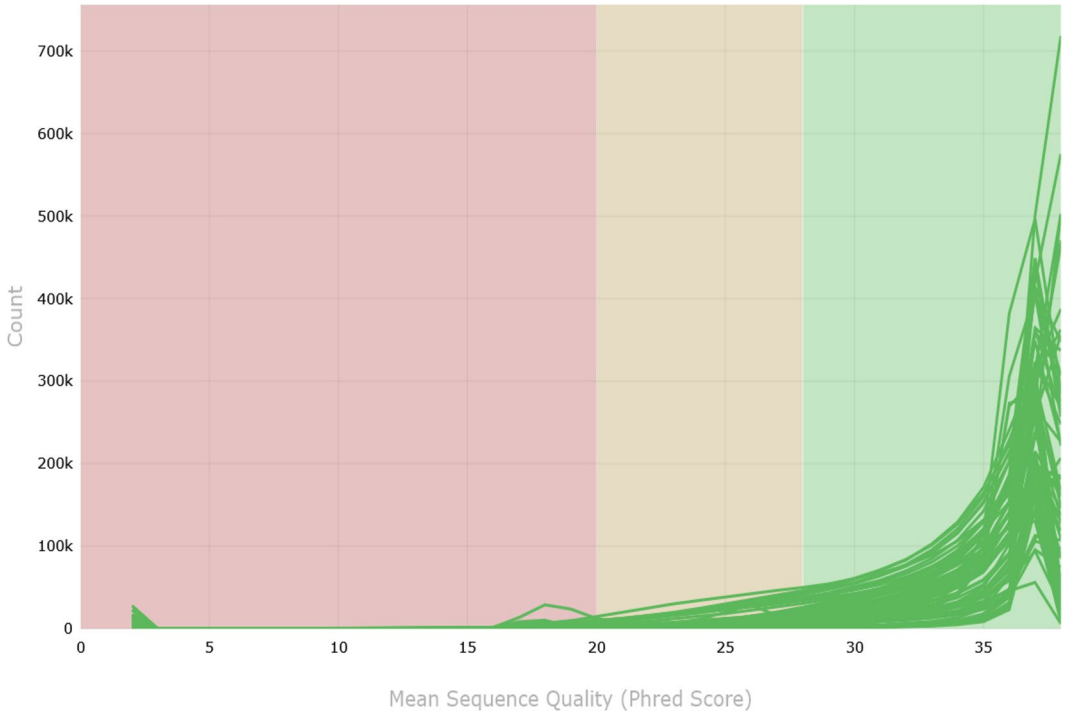

Fig. 5 Per Sequence Quality Scores.

pathways derived from functional annotation and bacterial species, total species, total genera, and total families from the ribosomally active community. Examining these curves indicates that the collected data sufficiently captures and represents the diversity within the samples, thereby ensuring the reliability of the sampling efforts.

In this context, it is essential to distinguish between rarefaction curves and species accumulation curves, as both play a crucial role in the technical validation of sampling efforts in metatranscriptomics studies, particularly when examining the saline-activated sludge microbiome. Rarefaction curves, which illustrate the cumulative number of observed functional features—such as metabolic pathways or ribosomally active species—relative to sequencing depth, offer a quantitative measure of sequencing adequacy. These curves facilitate the assessment of sampling completeness and help identify the point of sequencing saturation. On the other hand, species and metabolic pathway accumulation curves, which track the total number of species and pathways detected as a function of sample number, are instrumental in evaluating the thoroughness of sample collection and its capacity to capture the full breadth of community diversity across the sampling effort.

Benchmarking with existing scientific knowledge. Figure 2 illustrates the relative abundance of bacterial phyla across the 32 samples, providing a valuable benchmark for assessing the taxonomic alignment of the saline-activated sludge microbiome. It is well-established that the bacteriota plays a crucial role in driving removal efficiency in activated sludge systems, which has led to extensive study, particularly through 16S rRNA sequencing[23]. The resulting bacterial composition of the saline-activated sludge aligns with existing knowledge in activated sludge, comprising phyla such as *Pseudomonadota, Bacillota, Bacteroidota, Actinomycetota, Chloroflexota*[43,44]. Notably, *Pseudomonadota* emerges as the most abundant phylum within the saline-activated sludge, which is consistent with prior studies, due to its involvement in critical processes such as denitrification, organic matter degradation, and adaptation to fluctuating environmental conditions[45–47].

The *Thermodesulfobacteriota*, while not commonly abundant in many activated sludge systems, is notably enriched in this saline activated sludge bacteriota. The presence of *Thermodesulfobacteriota* suggests the importance of sulfur-reducing bacteria in the system, especially under saline conditions, where they may contribute to sulfur cycling and the reduction of sulfate to hydrogen sulfide[48]. These observations underscore the importance of studying the saline-activated sludge microbiome to better understand the unique microbial processes at play in these environments.

Usage Notes

This data can be used to explore several key aspects of microbiome functionality within wastewater treatment systems. Firstly, it allows for a comparative analysis between microbiomes in saline and domestic wastewater treatment plants, highlighting differences in microbial community structures and their functional roles across these distinct activated sludge ecosystems. Secondly, it facilitates longitudinal monitoring of the microbiome within a wastewater treatment plant, providing valuable insights into microbial dynamics and responses to environmental/process parameters over time. Finally, the data offers the opportunity to examine a specific Dutch wastewater treatment facility, contributing to a detailed understanding of microbiome characteristics and performance unique to the Netherlands, which can enhance local wastewater treatment practices and environmental management strategies.

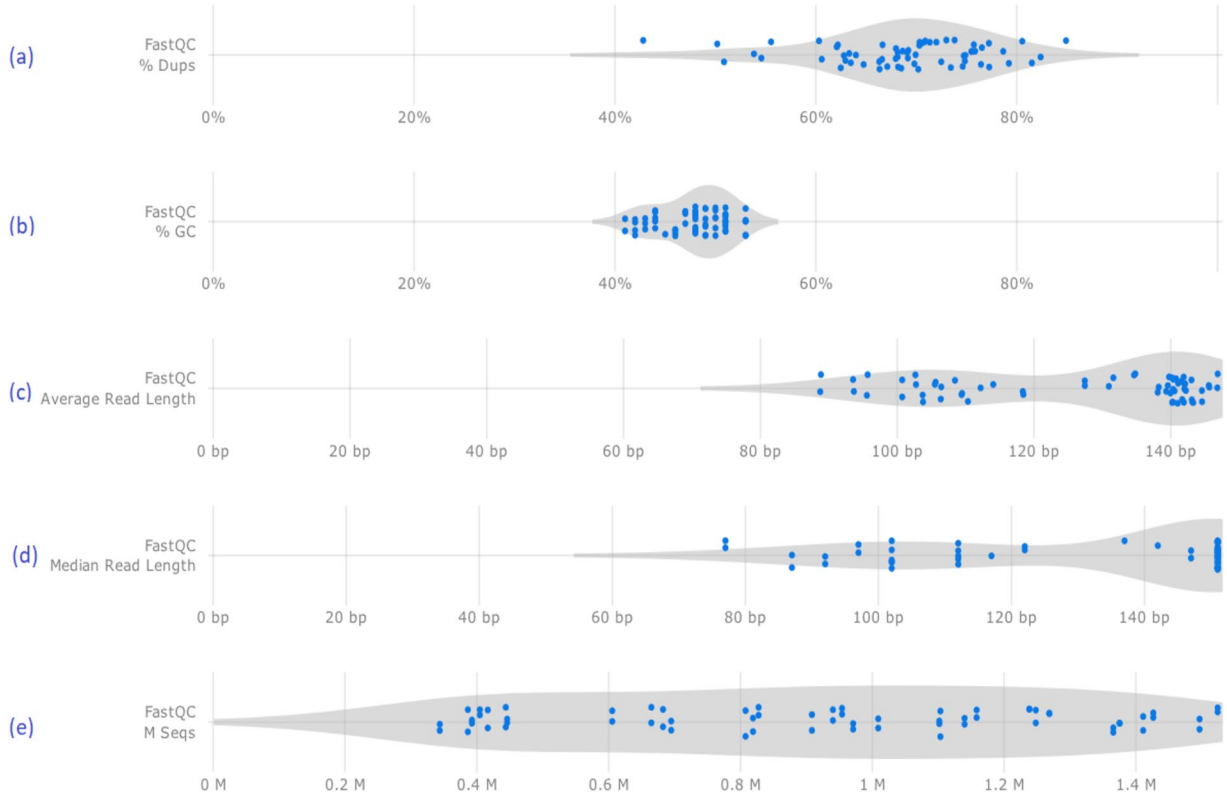

Fig. 6 Violin plot summarizing the general statistics of sequence data. (**a**) Distribution of percentage of duplicate reads among the 32 metatranscriptomes. (**b**) Distribution of average percentage of GC content among the 32 metatranscriptomes. (**c**) Distribution of the average read length in base pairs (bp) among the 32 metatranscriptomes. (**d**) Distribution of the median read length in base pairs (bp) among the 32 metatranscriptomes. (**e**) Distribution of the total sequences in millions (M) among the 32 metatranscriptomes.

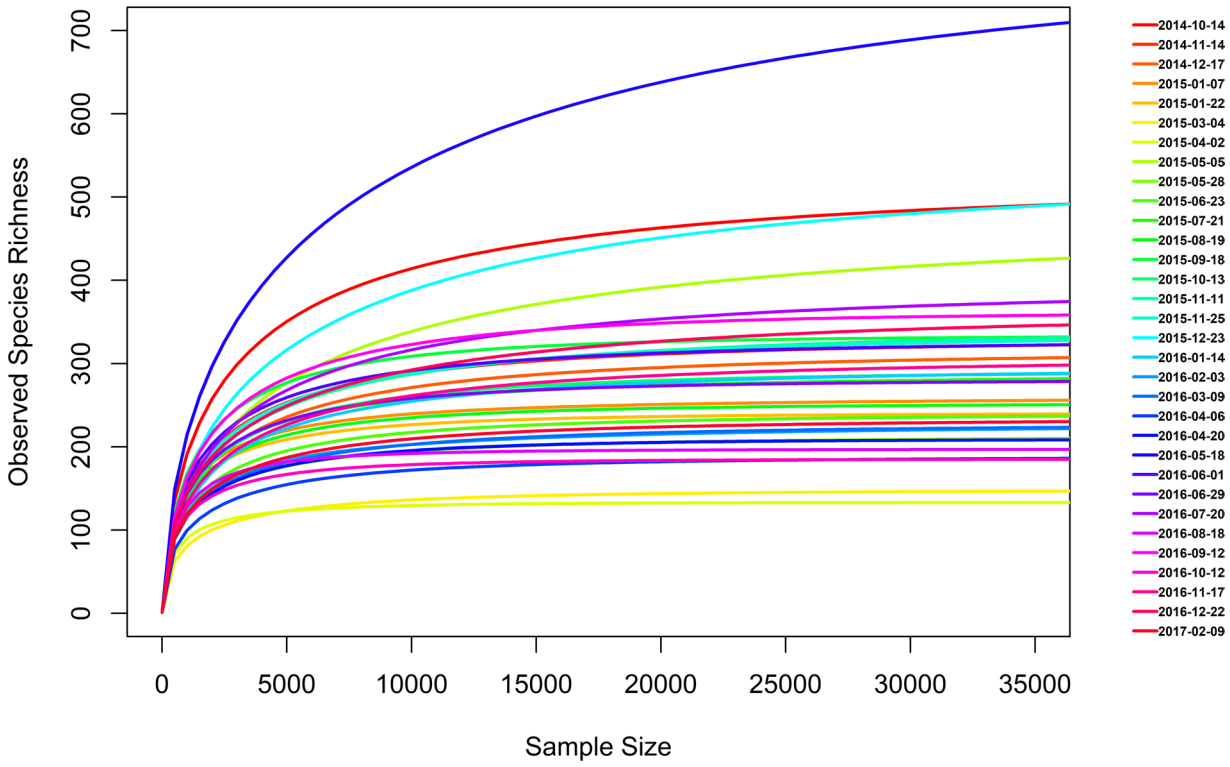

Fig. 7 Rarefaction curves illustrating observed species richness as a function of sequencing depth across 32 samples.

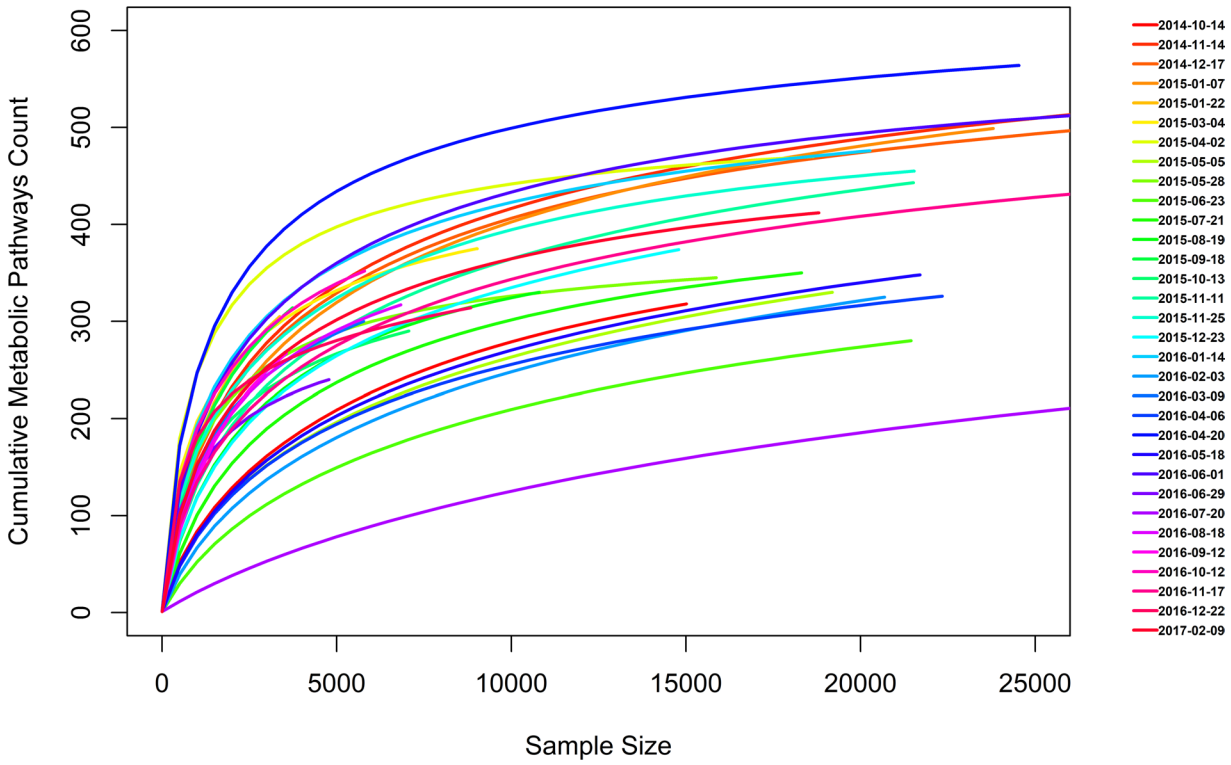

Fig. 8 Rarefaction curves showing the cumulative count of metabolic pathways as a function of sequencing depth across 32 samples.

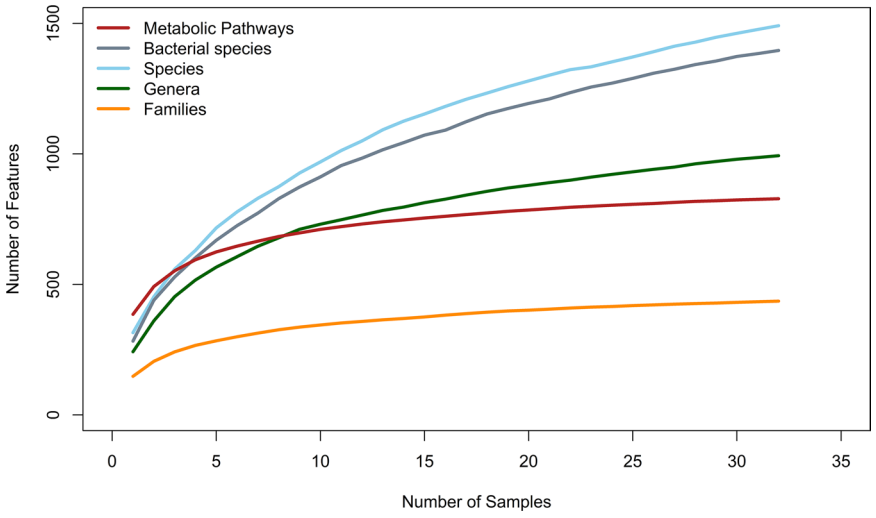

Fig. 9 Features accumulation curve.

Code availability

FastQC[27] (v0.12.1, https://github.com/s-andrews/FastQC) was used to conduct a quality assessment of the raw data. MultiQC[28] (v1.22.2, https://multiqc.info/) was used to aggregate and summarize results from FastQC. Groningen Microbiome Hub data analysis pipeline (https://github.com/GRONINGEN-MICROBIOME-CENTRE/GMH_MGS_pipeline) can be used for total community profiling. SAMSA2[35] (v.2.2.0, https://github.com/transcript/samsa2) pipeline was used to predict microbial functions encoded in the meta-transcriptomes. The data analysis was performed using SAMSA2 master analysis script modified to use UMCG HPC cluster (https://github.com/GRONINGEN-MICROBIOME-CENTRE/GMH_MGS_pipeline/blob/main/utils/metatranscriptomics/SAMSA2_master_script_RUG_habrok.sh).

Received: 7 October 2024; Accepted: 19 February 2025;
Published online: 26 February 2025

References

1. Wu, L. *et al*. Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nature Microbiology* **4**, 1183–1195, https://doi.org/10.1038/s41564-019-0426-5 (2019).
2. Sanz, J. L. & Köchling, T. Next-generation sequencing and waste/wastewater treatment: a comprehensive overview. *Reviews in Environmental Science and Bio/Technology* **18**, 635–680, https://doi.org/10.1007/s11157-019-09513-0 (2019).
3. Negara, M. A. P., Cornelissen, E., Geurkink, A. K., Euverink, G. J. W. & Jayawardhana, B. Next Generation Sequencing Analysis of Wastewater Treatment Plant Process via Support Vector Regression. *IFAC PapersOnLine* **52**, 37–42, https://doi.org/10.1016/j.ifacol.2019.11.006 (2019).
4. Mahajna, A., Dinkla, I. J. T., Euverink, G. J. W., Keesman, K. J. & Jayawardhana, B. Clean and Safe Drinking Water Systems via Metagenomics Data and Artificial Intelligence: State-of-the-Art and Future Perspective. *Frontiers in Microbiology* **13**, https://doi.org/10.3389/fmicb.2022.832452 (2022).
5. Yu, L. *et al*. Biodiversity, isolation and genome analysis of sulfamethazine-degrading bacteria using high-throughput analysis. *Bioprocess and Biosystems Engineering* **43**, 1521–1531, https://doi.org/10.1007/s00449-020-02345-1 (2020).
6. Liu, H. *et al*. Stable-isotope probing coupled with high-throughput sequencing reveals bacterial taxa capable of degrading aniline at three contaminated sites with contrasting pH. *The Science of the total environment* **771**, 144807, https://doi.org/10.1016/j.scitotenv.2020.144807 (2021).
7. Singleton, C. M. *et al*. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature communications* **12**, 2009, https://doi.org/10.1038/s41467-021-22203-2 (2021).
8. Ekholm, J. *et al*. Microbiome structure and function in parallel full-scale aerobic granular sludge and activated sludge processes. *Applied microbiology and biotechnology* **108**, 334, https://doi.org/10.1007/s00253-024-13165-8 (2024).
9. Orschler, L., Agrawal, S. & Lackner, S. Targeted metagenomics reveals extensive diversity of the denitrifying community in partial nitritation anammox and activated sludge systems. *Biotechnology and bioengineering* **118**, 433–441, https://doi.org/10.1002/bit.27581 (2021).
10. Sato, Y. *et al*. Transcriptome analysis of activated sludge microbiomes reveals an unexpected role of minority nitrifiers in carbon metabolism. *Communications Biology* **2**, 1–8, https://doi.org/10.1038/s42003-019-0418-2 (2019).
11. Zou, K. *et al*. Cyanophycin Granule Polypeptide: a Neglected High Value-Added Biopolymer, Synthesized in Activated Sludge on a Large Scale. *Applied and Environmental Microbiology* **88**, https://doi.org/10.1128/aem.00742-22 (2022).
12. Dueholm, M. K. D. *et al*. Genetic potential for exopolysaccharide synthesis in activated sludge bacteria uncovered by genome-resolved metagenomics. *Water research* **229**, 119485, https://doi.org/10.1016/j.watres.2022.119485 (2023).
13. Geng, S. *et al*. Effects of an external magnetic field on microbial functional genes and metabolism of activated sludge based on metagenomic sequencing. *Scientific reports* **10**, 8818, https://doi.org/10.1038/s41598-020-65795-3 (2020).
14. Wang, Y. *et al*. Genome-centric metagenomics reveals the host-driven dynamics and ecological role of CPR bacteria in an activated sludge system. *Microbiome* **11**, 56, https://doi.org/10.1186/s40168-023-01494-1 (2023).
15. Zhang, J. *et al*. Regional discrepancy of microbial community structure in activated sludge system from Chinese WWTPs based on high-throughput 16S rDNA sequencing. *Science of the Total Environment* **818**, https://doi.org/10.1016/j.scitotenv.2021.151751 (2022).
16. Shi, X. *et al*. Seasonal effects on pilot-scale high-concentration activated sludge systems in cold regions. *Journal of Water Process Engineering* **52**, https://doi.org/10.1016/j.jwpe.2023.103575 (2023).
17. Zhao, F. *et al*. Correlations among Antibiotic Resistance Genes, Mobile Genetic Elements and Microbial Communities in Municipal Sewage Treatment Plants Revealed by High-Throughput Sequencing. *International journal of environmental research and public health* **20**, https://doi.org/10.3390/ijerph20043593 (2023).
18. Yadav, S. & Kapley, A. Exploration of activated sludge resistome using metagenomics. *The Science of the total environment* **692**, 1155–1164, https://doi.org/10.1016/j.scitotenv.2019.07.267 (2019).
19. Huang, Y., Zou, K., Qing, T., Feng, B. & Zhang, P. Metagenomics and metatranscriptomics analyses of antibiotic synthesis in activated sludge. *Environmental Research* **213**, https://doi.org/10.1016/j.envres.2022.113741 (2022).
20. Sun, Y., Clarke, B., Clarke, J. & Li, X. Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning. *Water research* **202**, 117384, https://doi.org/10.1016/j.watres.2021.117384 (2021).
21. Gruber, W. *et al*. Linking seasonal N2O emissions and nitrification failures to microbial dynamics in a SBR wastewater treatment plant. *Water Research X* **11**, https://doi.org/10.1016/j.wroa.2021.100098 (2021).
22. Liang, Z., Yi, J., Gu, Q. & Dai, X. Metagenomics reveals a full-scale modified integrated fixed-film activated sludge process: Enhanced nitrogen removal and reduced sludge production. *The Science of the total environment* **841**, 156666, https://doi.org/10.1016/j.scitotenv.2022.156666 (2022).
23. Dueholm, M. K. D. *et al*. MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nature Communications* **13**, 1908, https://doi.org/10.1038/s41467-022-29438-7 (2022).
24. *North Water Saline Wastewater Treatment Plant* https://northwater.nl/wp-content/uploads/2018/10/Leaflet-ZAWZI-English.pdf (2018).
25. de Boks, P. A., de Wit, M. & Menkveld, W. *Purification of salty wastewater from companies near Delfzijl* https://edepot.wur.nl/341746 (2009).
26. Mahajna, A. All process data of Oosterhorn SWWTP case study. *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.27073612.v2 (2024).
27. FastQC: A quality control analysis tool for high throughput sequencing data (2023).
28. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)* **32**, 3047–3048, https://doi.org/10.1093/bioinformatics/btw354 (2016).
29. Abueg, L. A. L. *et al*. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic acids research* **52**, W83–W94, https://doi.org/10.1093/nar/gkae410 (2024).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
31. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* **15**, R46, https://doi.org/10.1186/gb-2014-15-3-r46 (2014).
32. Liu, Y., Ghaffari, M. H., Ma, T. & Tu, Y. Impact of database choice and confidence score on the performance of taxonomic classification using Kraken2. *aBIOTECH: An International Journal on Agricultural Biotechnology*, 1–11 https://doi.org/10.1007/s42994-024-00178-0 (2024).
33. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733–745, https://doi.org/10.1093/nar/gkv1189 (2016).
34. Jennifer, L., Florian, P. B., Peter, T. & Steven, L. S. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**, e104, https://doi.org/10.7717/peerj-cs.104 (2017).
35. Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I. & Lemay, D. G. SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC bioinformatics* **19**, 175, https://doi.org/10.1186/s12859-018-2189-z (2018).
36. Jiajie, Z., Kassian, K., Tomáš, F. & Alexandros, S. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620, https://doi.org/10.1093/bioinformatics/btt593 (2014).

37. Evguenia, K., Laurent, N. & Hélène, T. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217, https://doi.org/10.1093/bioinformatics/bts611 (2012).
38. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**, D206–D214, https://doi.org/10.1093/nar/gkt1226 (2014).
39. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods: Techniques for life scientists and chemists* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
40. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602, https://doi.org/10.1038/s41586-023-06583-7 (2023).
41. Mahajna, A. *NCBI BioProject.* http://identifiers.org/ncbi/BioProject:PRJNA1122484 (2024).
42. Mahajna, A. *Sequence Read Archive.* http://identifiers.org/ncbi/insdc.sra:SRP513109 (2024).
43. Begmatov, S. *et al.* The structure of microbial communities of activated sludge of large-scale wastewater treatment plants in the city of Moscow. *Sci Rep* **12**, 3458, https://doi.org/10.1038/s41598-022-07132-4 (2022).
44. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol* **4**, 1183–1195, https://doi.org/10.1038/s41564-019-0426-5 (2019).
45. Wasmund, K., Singleton, C., Dahl Dueholm, M. K., Wagner, M. & Nielsen, P. H. The predicted secreted proteome of activated sludge microorganisms indicates distinct nutrient niches. *mSystems* **9**, e0030124, https://doi.org/10.1128/msystems.00301-24 (2024).
46. Kim, S. J. *et al.* Genomic and metatranscriptomic analyses of carbon remineralization in an Antarctic polynya. *Microbiome* **7**, 29, https://doi.org/10.1186/s40168-019-0643-4 (2019).
47. Oztas Gulmus, E. & Gormez, A. Characterization and biotechnological application of protease from thermophilic Thermomonas haemolytica. *Arch Microbiol* **202**, 153–159, https://doi.org/10.1007/s00203-019-01728-7 (2020).
48. Zhang, H., Song, S., Jia, Y., Wu, D. & Lu, H. Stress-responses of activated sludge and anaerobic sulfate-reducing bacteria sludge under long-term ciprofloxacin exposure. *Water Res* **164**, 114964, https://doi.org/10.1016/j.watres.2019.114964 (2019).

Acknowledgements

This work was performed in the cooperation framework of Wetsus, European Centre of excellence for sustainable water technology (www.wetsus.nl). Wetsus is co-funded by the Dutch Ministry of Economic Affairs and Climate Policy, the European Union Regional Development Fund, the Province of Fryslân and the Northern Netherlands Provinces. The authors would like to thank the participants of the research theme "Genomics Based Water Quality Monitoring" for the fruitful discussions and their financial support. We extend our sincere gratitude to Inez Dinkla, Theme Coordinator for the "Genomics-Based Water Quality Monitoring" research group at Wetsus, for her invaluable support throughout all stages of this research. We also wish to acknowledge BioClear Earth bv, a company participant within Wetsus, for their contributions to the sequencing efforts and thank Elsemiek Croese from BioClear Earth for her technical assistance with sample preparation and sequencing. We thank North Water for their support in the sampling effort. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to high performance computing clusters.

Author contributions

A.M. drafted the manuscript and conducted downstream analysis and visualization of microbiome data. B.G. designed the sampling campaign, collected data, provided oversight of the samples' pre-treatment and sequencing, and provided data stewardship. R.G. designed and implemented the metatranscriptomics pipeline. K.J.K., G.J.W.E., B.J. conceived, coordinated, and supported the study. All authors critically revised and approved the manuscript.

Competing interests

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.

Additional information

Correspondence and requests for materials should be addressed to A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025