Metaproteomics, metagenomics and 16S rRNA sequencing provide different perspectives on the aerobic granular sludge microbiome

Kleikamp, Hugo B.C.; Grouzdev, Denis; Schaasberg, Pim; van Valderen, Ramon; van der Zwaan, Ramon; Wijgaart, Roel van de; Lin, Yuemei; Abbas, Ben; Pronk, Mario; van Loosdrecht, Mark C.M.

DOI
10.1016/j.watres.2023.120700

Publication date
2023

Document Version
Final published version

Published in
Water Research

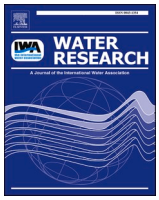
Citation (APA)
Kleikamp, H. B. C., Grouzdev, D., Schaasberg, P., van Valderen, R., van der Zwaan, R., Wijgaart, R. V. D., Lin, Y., Abbas, B., Pronk, M., van Loosdrecht, M. C. M., & Pabst, M. (2023). Metaproteomics, metagenomics and 16S rRNA sequencing provide different perspectives on the aerobic granular sludge microbiome. *Water Research*, *246*, Article 120700. https://doi.org/10.1016/j.watres.2023.120700

Water Research 246 (2023) 120700

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres



Metaproteomics, metagenomics and 16S rRNA sequencing provide different perspectives on the aerobic granular sludge microbiome

Hugo B.C. Kleikamp [a,1,*], Denis Grouzdev [b], Pim Schaasberg [a], Ramon van Valderen [a], Ramon van der Zwaan [a], Roel van de Wijgaart [a], Yuemei Lin [a], Ben Abbas [a], Mario Pronk [a], Mark C.M. van Loosdrecht [a], Martin Pabst [a,*]

[a] Department of Biotechnology, Delft University of Technology, Delft, the Netherlands
[b] SciBear OU, Tallinn, Estonia

ARTICLE INFO

Keywords:
Metaproteomics
metagenomics
16S rRNA amplicon sequencing
Aerobic granular sludge
Wastewater treatment

ABSTRACT

The tremendous progress in sequencing technologies has made DNA sequencing routine for microbiome studies. Additionally, advances in mass spectrometric techniques have extended conventional proteomics into the field of microbial ecology. However, systematic studies that provide a better understanding of the complementary nature of these 'omics' approaches, particularly for complex environments such as wastewater treatment sludge, are urgently needed.

Here, we describe a comparative metaomics study on aerobic granular sludge from three different wastewater treatment plants. For this, we employed metaproteomics, whole metagenome, and 16S rRNA amplicon sequencing to study the same granule material with uniform size. We furthermore compare the taxonomic profiles using the Genome Taxonomy Database (GTDB) to enhance the comparability between the different approaches. Though the major taxonomies were consistently identified in the different aerobic granular sludge samples, the taxonomic composition obtained by the different omics techniques varied significantly at the lower taxonomic levels, which impacts the interpretation of the nutrient removal processes. Nevertheless, as demonstrated by metaproteomics, the genera that were consistently identified in all techniques cover the majority of the protein biomass. The established metaomics data and the contig classification pipeline are publicly available, which provides a valuable resource for further studies on metabolic processes in aerobic granular sludge.

1. Introduction

Microbial communities play a central role in the global biogeochemical cycles, and their close association with humans has a direct impact on health and disease (Cho and Blaser, 2012; Falkowski et al., 2008; Integrative et al., 2019; Rousk and Bengtson, 2014; Turnbaugh et al., 2007). Moreover, microbial communities are increasingly used in biotechnology and engineering. For example, microbes are employed to degrade and remove pollutants from wastewater and soils, or microbes produce novel materials, greener chemicals, or energy to support a more sustainable society (Angenent et al., 2004; Balcom et al., 2016; Lovley, 2017; Rabaey and Verstraete, 2005; Tawalbeh et al., 2020; Temudo et al., 2008). Of more recent interest are also synthetic and engineered microbial communities. However, the complex nature of microbial interactions hampers efforts to engineer specific functions into such consortia (Kehe et al., 2019; Lawson, 2021). Therefore, systems biology approaches that provide molecular-level information from complex microbial communities become increasingly important in biotechnology

Abbreviations: MG, Metagenomics (whole metagenome sequencing); MP, Metaproteomics; 16S, 16S rRNA gene sequencing; AGS, Aerobic granular sludge; DXP, Dinxperlo, plant 1 (wastewater treatment plant, the Netherlands); GW, Garmerwolde, plant 2 (wastewater treatment plant, the Netherlands); SP, Simpelveld, plant 3 (wastewater treatment plant, the Netherlands); ASV, Amplicon sequence variant; PSM, Peptide-to-spectrum match; MAG, Metagenome assembled genome; GTDB, genome taxonomy database; COG, Cluster of orthologous groups of proteins; PAO, Polyphosphate-accumulating organisms; GAO, Glycogen-accumulating organisms; AOB, Ammonium-oxidizing bacteria; AOA, Ammonium-oxidizing archaea; NOB, Nitrite-oxidizing bacteria; NR, Nitrate-reducing organisms; EPS, Extracellular polymeric substances.

* Corresponding authors.
E-mail addresses: hugo.kleikamp@uantwerpen.be (H.B.C. Kleikamp), m.pabst@tudelft.nl (M. Pabst).
[1] Present address: Department of Biology, University of Antwerp, Antwerp, Belgium.

https://doi.org/10.1016/j.watres.2023.120700
Received 14 June 2023; Received in revised form 29 September 2023; Accepted 4 October 2023
Available online 6 October 2023

and microbial ecology.

The emergence of next-generation sequencing (NGS) technologies finally enabled large-scale genomic studies of microbial communities directly from their natural environments. The simplest of these approaches is 16S rRNA gene sequencing. 16S rRNA genes are highly conserved between different bacteria and archaea and, thus, are widely used to perform taxonomic profiling of environmental communities (Ali et al., 2019; de Sousa Rollemberg et al., 2019; Ramos et al., 2015; Wu et al., 2019; Zhang et al., 2012; Zhou and Sun, 2020). However, this approach suffers from variable 16S rRNA gene copy numbers (Louca et al., 2018; Starke et al., 2021; Stoddard et al., 2015) and primer efficiencies across microbes (Albertsen et al., 2015; Brown et al., 2015). Furthermore, metabolic functions are only inferred from prior taxonomic knowledge and, thus, remain purely predictive (Kyte and Doolittle, 1982; Morrissey et al., 2016). Alternatively, whole metagenome sequencing (commonly referred to as metagenomics) aims to sequence the genomes of all community members. Following assembly and binning into metagenome-assembled genomes (MAGs), the genomes can achieve strain-level resolution and provide the metabolic potential of individual community members (Jansson and Hofmockel, 2018; Ranjan et al., 2016; Rubio-Rincón et al., 2019). However, the sequences obtained through metagenomic sequencing may not only encompass the active microbial population, but may also cover free DNA and DNA from dead and dormant microbes (Quince et al., 2017). Advancements in RNA sequencing moreover enabled to measure the actively expressed genes in microbial communities. This approach, known as metatranscriptomics, however still faces challenges, such as obtaining high-quality RNA from biological samples and the short lifespan of mRNA, which hinders the detection of rapid or short-lived responses (Bashiardes et al., 2016). Advances in high-resolution mass spectrometry and the increased ease of constructing proteome sequence databases furthermore enabled deep metaproteomic studies on complete microbial communities (Hagen et al., 2017; Heyer et al., 2015; Kleiner et al., 2017; Muth et al., 2018; Püttker et al., 2015; Wilmes et al., 2015, 2008; Zorz et al., 2019). Most importantly, because metaproteomics measures the gene products (i.e., proteins), it provides a complementary view on the microbial community. The microbial composition obtained by metaproteomics correlates to the amount of protein biomass produced per microbe (Kleikamp et al., 2021; Kleiner et al., 2017). Therefore, metaproteomic data resemble more closely the metabolic capacity of individual community members (Blakeley-Ruiz et al., 2019; Kleiner, 2019; Salvato et al., 2021). Moreover, metaproteomics allows to measure regulatory events such as protein modifications, which cannot be obtained from genomic information alone (den Ridder et al., 2020; Li et al., 2014). However, in contrast to DNA, proteins cannot be amplified prior to analysis, and peptide sequencing is performed consecutively (or only at low multiplexing level) rather than in parallel. Therefore, the depth of information that can be obtained by metaproteomics depends on the taxonomic complexity and the mass spectrometric effort taken to sequence the sample (Hagen et al., 2017; Narayanasamy et al., 2015; Wilmes et al., 2015). The dependency of metaproteomic performance on community complexity has been investigated more in detail by Lohmann and co-workers also only recently (Lohmann et al., 2020). Information obtained from DNA and rRNA-based experiments, have often been found to contradict staining experiments or measured metabolic conversions (Azizan et al., 2020; de Sousa Rollemberg et al., 2019; Welles et al., 2015). This highlights the importance of employing additional (complementary) approaches–such as metaproteomics–when characterizing microbial communities.

Nevertheless, the lack of standardization within the omics field makes comparison of different experiments challenging, even if studies used the same omics approach (Balvočiūtė and Huson, 2017; Sczyrba et al., 2017; Van Den Bossche et al., 2021). For example, metagenomics experiments are commonly employed to construct protein sequence databases for metaproteomics studies to enable deep sequence coverage and high taxonomic and functional resolution. A comprehensive taxonomic classification of the metagenomic data, however, relies on accurate and complete reference sequence databases. Consequently, a potential source of large variation and inaccuracy already derives from the reference databases used to taxonomically classify the metagenomic sequences. Different reference databases vary substantially in taxonomic coverage, sequence content, and the nomenclature and employed phylogenies. Modern phylogenetic placement tools use a range of methods such as 16S similarity, average amino acid identity and average nucleotide identity (Konstantinidis and Tiedje, 2005; Yarza et al., 2014). The NCBI taxonomy, which is used for RefSeq and UniProtKB, employs a mixture of historical taxonomies and modern placement methods and it lacks a rank normalization. This results in lineages with gaps in taxonomic annotations (further referred to as ‘gapped lineages’) (Federhen, 2012; Schoch et al., 2020). In addition, NCBI taxonomies contain clusters that group uncultured organisms (further referred to as ‘dump taxa’) (Hugenholtz et al., 2016). Thus, the NCBI taxonomy is often not consistent with respect to true evolutionary relationships. Many taxa circumscribe polyphyletic groupings and there is an uneven application of ranks across the phylogenetic tree (Abbott and Janda, 2006; McDonald et al., 2012; Parks et al., 2018). Standardized reference sequence databases with accurate taxonomies are therefore of utmost importance for accurately describing microbial diversity, enabling comparison between experiments, and communicating scientific data (Godfray, 2002; Parks et al., 2018). The recently established genome taxonomy database (GTDB) addresses these issues by using a set of conserved proteins and employing a placement method that normalizes taxonomic ranks based on relative evolutionary divergence (Chaumeil et al., 2020; Parks et al., 2020, Parks et al., 2022; 2018). The GTDB taxonomy offers an objective, phylogenetically-consistent classification of prokaryotic species, and therefore enables a more accurate description of the taxonomic and metabolic diversity of a microbial community (Parks et al., 2018). The Genome Taxonomy Database Toolkit (GTDB-Tk) supports the classification of draft bacterial and archaeal genomes (Chaumeil et al., 2020). However, GTDB-Tk was developed for genome assemblies or metagenome-assembled genomes constructed by clustering related contigs into bins (Lin et al., 2021; Sedlar et al., 2017). The binning procedure, however, leaves a substantial fraction of unbinned sequences for complex metagenomes (Sczyrba et al., 2017). Consequently, assembled genomes often provide a substantially less complete sequence reference database compared to the alternative reads- or contig-based databases (Chen et al., 2020; Jouffret et al., 2021; May et al., 2016; Olson et al., 2019; Tanca et al., 2016), which is a major limitation for metaproteomic studies. For that reason, database construction and taxonomic classification have been frequently performed on contigs or scaffolds, e.g. as demonstrated by the contig annotation tool (CAT) (von Meijenfeldt et al., 2019).

From the many applications in industrial biotechnology, microbial water treatment is perhaps one of the fastest-growing areas. For example, the activated sludge process is the most widely employed biological wastewater treatment process to purify wastewater in developed areas (Orhon et al., 2009; van Loosdrecht and Brdjanovic, 2014). Large-scale 16S rRNA sequencing efforts on activated sludge established the wastewater microbiome specific database termed ‘MiDAS’ (Microbial Database for Activated Sludge). The consortium created a global map of microbes present in activated sludge systems, with the aim of linking organisms to nutrient-removal functions [38–40]. An advancement of this process – known as aerobic granular sludge (AGS) technology – has the advantage of operating with reduced space and energy requirements (de Sousa Rollemberg et al., 2019; Pronk et al., 2015; Świątczak and Cydzik-Kwiatkowska, 2018). In AGS, the microbes form dense granules following the production of extracellular polymeric substances (Adav et al., 2009; Liang et al., 2019; Panchavinin et al., 2019). Consequently, the granules allow a higher settling speed and biomass density. In microbial wastewater treatment, several synergistic roles for nutrient removal have been identified that include polyphosphate-accumulating organisms (PAO), glycogen-accumulating

organisms (GAO), nitrite-oxidizing bacteria (NOB), ammonia-oxidizing bacteria (AOB), and nitrate reducers (NR) (Szabó et al., 2017; Weissbrodt et al., 2013, 2014). Although microbial wastewater treatment has a long history, the exact molecular-level processes and the organisms that are involved in nutrient removal are still poorly understood (Ali et al., 2019; Leventhal et al., 2018). Therefore, determining the taxonomic composition of the core microbiome and the expressed metabolic functions are important in optimizing purification processes and developing advanced purification strategies.

Here, we provide the first comparative metaomics study on aerobic granular sludge microbiome, which was sampled from three different wastewater treatment plants. We systematically compare the taxonomic and metabolic profiles obtained by the different omics approaches as well as reference sequence databases. The established data demonstrate the different perspectives that can be obtained on the aerobic granular sludge microbiome, which provides a valuable resource for future studies on the nutrient removal processes.

2. Materials and methods

2.1. Sampling of aerobic granular sludge

Aerobic granular sludge (AGS) was collected from three different full-scale AGS wastewater treatment plants in the Netherlands: Dinxperlo (DX, plant 1), Garmerwolde (GW, plant 2) and Simpelveld (SP, plant 3). Each plant performed stable operation with simultaneous denitrification and phosphorus removal. AGS granules were sieved to select a size fraction with a diameter of approximately 2.0 mm. Granules were stored at −80 °C until further processed

2.2. Protein extraction and proteolytic digestion

The collected granules were freeze-dried and ground with a mortar and pestle. Two hundred milligrams of acid washed glass beads (150–212 μm) and 350 μL of both TEAB and B-PER buffer were added to approximately 5 mg starting material. Bead beating was performed for 20 s (× 3) with a 30 s pause between cycles. Samples were centrifuged and freeze/thaw cycles (× 3) were performed by freezing the sample at −80 °C and subsequently thawing at 95 °C in a water bath. The samples were centrifuged, and the supernatant was collected. Protein precipitation was performed by adding TCA at a ratio of TCA to supernatant of 1:4. The samples were incubated at 4 °C for 10 min. and then centrifuged at 14,000 r.p.m. for 5 min. The pellets were washed with 200 μL ice-cold acetone. The protein pellets were reconstituted in 250 μL 6 M urea and the protein extracts were then reduced with 10 mM dithiothreitol (DTT) for 60 min. at 37 °C. Next, the samples were alkylated with 20 mM iodoacetamide (IAA) and incubated in the dark at room temperature for 30 min. Thereafter, the samples were diluted with 200 mM ammonium bicarbonate (AmBiC) to <1 M urea. Finally, sequencing-grade trypsin was added (Promega) at an approximate enzyme to protein ratio of 1:50 and incubated at 37 °C overnight. The obtained peptides were purified by solid-phase extraction using Oasis HLB solid-phase extraction well plates (Waters) according to the protocol provided by the manufacturer. Purified peptide fractions were then dried in a SpeedVac concentrator, reconstituted in aqueous 0.1 % TFA and separated (according to the instructions supplied by the manufacturer) into 8 fractions using the Pierce high pH reversed-phase fractionation kit (Thermo Scientific). For plants 2 (DX) and 3 (GW) the fractions 2 + 6, 3 + 7, 4 + 8 were combined. The obtained samples were dried in a SpeedVac concentrator and dissolved in water containing 3 % acetonitrile and 0.1 % formic acid, resulting in 8 fractions for plant 1 (DX), 4 fractions for plant 2 (GW) and 3 (SP). The approximate concentration of the protein digest was determined using a NanoDrop micro-volume spectrophotometer.

2.3. Shotgun metaproteomic analysis

Briefly, the prepared fractions were analyzed by injecting approx. 300 ng proteolytic digest using a one-dimensional shotgun proteomic approach on a nano-liquid-chromatography system consisting of an EASY nano-LC 1200 equipped with an Acclaim PepMap RSLC RP C18 separation column (50 μm × 150 mm, 2 μm and 100 Å) coupled to a QE Plus Orbitrap mass spectrometer (Thermo Scientific, Germany). The flow rate was maintained at 350 nL/min using as solvent A water containing 0.1 % formic acid, and as solvent B 80 % acetonitrile in water and 0.1 % formic acid. The Orbitrap was operated in data-dependent acquisition mode acquiring peptide signals at 70 K resolution and a max IT of 100 ms, where the top 10 precursor ions were isolated by a 2.0 *m/z* window with an 0.1 *m/z* isolation offset, and fragmented at an NCE of 28. The AGC target was set to 2e5 at a max. IT of 75 ms and 17.5 K resolution. Mass peaks with unassigned charge state, singly, 7 and >7, were excluded from fragmentation. For the prepared fractions from plants 2 (GW) and 3 (SP) analysis in duplicates was performed using a linear gradient from 5 % to 28 % solvent B for 115 min and finally to 55 % B for additional 60 min. The individual fractions obtained from plant 1 were analysed by single injections using a short linear gradient from 6 % to 26 % solvent B for 45 min and finally to 50 % B over additional 10 min.

2.4. Processing of metaproteomic raw data

Mass spectrometric raw data (obtained from the fractions of each plant) were combined and analysed using PEAKS StudioX (Bioinformatics Solutions Inc., Canada) by either database searching against the metagenomic-constructed databases from predicted ORFs, or by *de novo* sequencing as quality control and to estimate the percentage of eukaryotic sequences (Kleikamp et al., 2021; Pabst et al., 2021). Redundant sequences in the constructed databases were removed by employing a local installation of CD-HIT, by clustering ORFs at 100 % identity (Li and Godzik, 2006). Database searching was performed by including cRAP protein sequences (https://www.thegpm.org/crap/), setting carbamidomethylation (C) as fixed and oxidation (M) and deamidation (N/Q) as variable modifications, allowing up to 2 missed cleavages and 2 variable modifications per peptide. Peptide-spectrum matches were filtered for 1 % false discovery rate. Protein identifications with ≥2 unique peptides were considered as significant. Taxonomic annotation of database-matched peptide sequences was based on the taxonomic classification obtained for the contigs (see below for taxonomic classification of metagenomics data). Metabolic annotation of ORFs using BlastKOALA (Kanehisa et al., 2016) was performed to obtain KEGG orthologies. WEBMGA (Wu et al., 2011) was used to annotate Clusters of Orthologous Groups (COGs) and protein families, PFAMs and the complementary TIGRFAM terms. DIAMOND v2.11 (Buchfink et al., 2015) was employed to annotate ORFs with UniprotKB genes.

2.5. DNA extraction and sequencing

DNA extraction was done using a DNeasy UltraClean Microbial Kit (Qiagen, Germany). The extracted DNA was quantified with a Qubit fluorometer. 16S rRNA gene amplification (by amplifying V3–V4 regions with 341F, 806R primers) and sequencing, and whole metagenome sequencing was performed on an Illumina NovaSeq platform with paired-end reads (Novogene Co. Ltd., China).

2.6. Processing of 16S rRNA raw sequencing data

Amplicons were sequenced on an Illumina paired-end platform to generate 250 bp paired-end raw reads. Subsequently, FLASH (V1.2.7) (Magoč and Salzberg, 2011) was used to merge the paired-end reads into raw tags. The raw tags were then subjected to quality filtering, using the

Qiime (V1.7.0, http://qiime.org/scripts/split_libraries_fastq.html) quality control process (Bokulich et al., 2013; Caporaso et al., 2010). Any chimeric sequences present in the clean tags were identified and removed using the UCHIME algorithm and the "Gold reference database" (Edgar et al., 2011). From cleaned reads amplicon sequence variants (ASVs) were selected with Usearchv11 command -unoise3 (Edgar, 2017). The data were padded with additional samples from each water treatment plant to improve ASV selection (data not shown). Taxonomic annotation was performed using QIIME2 with trained V3–V4 classifiers (Bolyen et al., 2019). 16S rRNA sequences were furthermore annotated with GTDB ssu rRNA (small subunit ribosomal RNA) sequences, Midas 3.7 flASVs, and a SILVA NR99 (v138) pre-trained V3–V4 classifiers (Bokulich et al., 2018).

2.7. Processing of whole metagenome sequencing raw data

Raw reads were quality checked with FastQC v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and low-quality reads were trimmed using Trimmomatic v0.39 with the default settings for pair-end reads (Bolger et al., 2014). Subsequently, reads were assembled with metaSPAdes v3.14.0 using default settings (Nurk et al., 2017). Prodigal v2.6.3 was employed to identify open reading frames (ORFs) (Hyatt et al., 2010). DIAMOND v2.11 was used to align identified ORFs to GTDB r202, UniProtKB (release 2021/03), Swiss-Prot UniRef100, UniRef90, UniRef50, and NCBI RefSeq protein and RefSeq protein non-redundant release 205 (Buchfink et al., 2015), using parameters -fast -top 10 -e 0.001 (and otherwise default parameters). A contig-level taxonomic classification was furthermore achieved based on the 'CAT' approach, as published by von Meijenfeldt et al. (2019). Briefly, the taxonomy of each ORF was determined by lowest common ancestor analysis of the top Diamond hits followed by constructing a consensus lineage for each contig from the classified ORFs. Adjustments compared to the original CAT approach were made with the objective to maximize genus level annotations of the dominant taxonomies. The LCA parameter selection was guided by the 16S rRNA amplicon sequencing data. A detailed description of the enhanced LCA approach can be found in the supplementary information material chapter 1 and SI Figs. 1–4. The developed Python codes for preprocessing GTDB sequences for the use with Diamond and for performing the 'protein LCA' are available via https://github.com/hbckleikamp/GTDB2DIAMOND. Python codes for reformatting sequences for the use with QIIME are available via: https://github.com/hbckleikamp/GTDB2QIIME. The metagenome coverage was estimated using Bowtie 2 v2.3.5.1 and QualiMap 2 v2.2.2 (Langmead and Salzberg, 2012; Okonechnikov et al., 2016) where the reads were first mapped to individual scaffolds using Bowtie and the obtained BAM file was analysed using QualiMap. The average depth of sequencing coverage was determined according to LN/G (L= length of read, N = number of reads and G = genome length) (Sims et al., 2014),

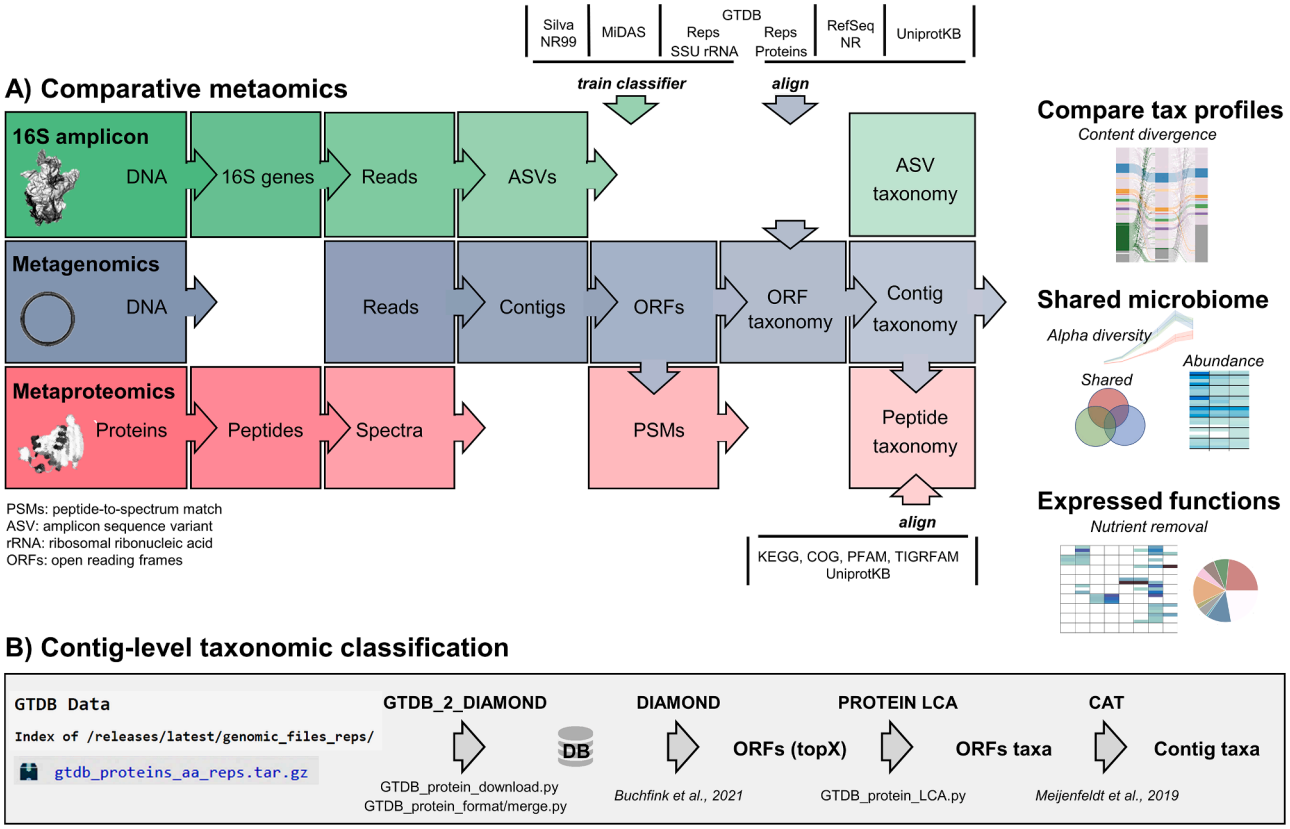


Fig. 1. A. The graph outlines the multi-omics approach used to characterize the aerobic granular sludge microbiome of three wastewater treatment plants. The same uniform 2 mm granule material was subjected to (i) metaproteomics (shown in red), (ii) whole metagenome sequencing (shown in blue) and (iii) 16S rRNA amplicon sequencing (shown in green). In whole metagenome sequencing, the reads were assembled into contigs, and the identified ORFs were aligned to reference sequence databases for taxonomic classification. Shotgun metaproteomics data were analyzed using a database containing the identified ORFs from the metagenomics experiments. For 16S rRNA gene sequencing, the amplicon sequencing variants (ASVs) were determined and compared to small subunit ribosomal RNA sequence databases for taxonomic classification. Additionally, a range of different reference sequence databases were used for taxonomic classification. The obtained taxonomic profiles and nutrient-removal pathways were compared between the different approaches and wastewater treatment plants. Fig. 1B) The scheme outlines the contig-based taxonomic classification using various reference sequence database, illustrated for the genome taxonomy database (GTDB). Reference sequences (protein reps) were downloaded from https://gtdb.ecogenomic.org/downloads/, merged, and reformatted to be compatible with the sequence aligner Diamond. A consensus lineage was determined based on the lowest common ancestor (LCA). Finally, the contig-level lineage was determined by employing a modified version of the CAT tool algorithm.

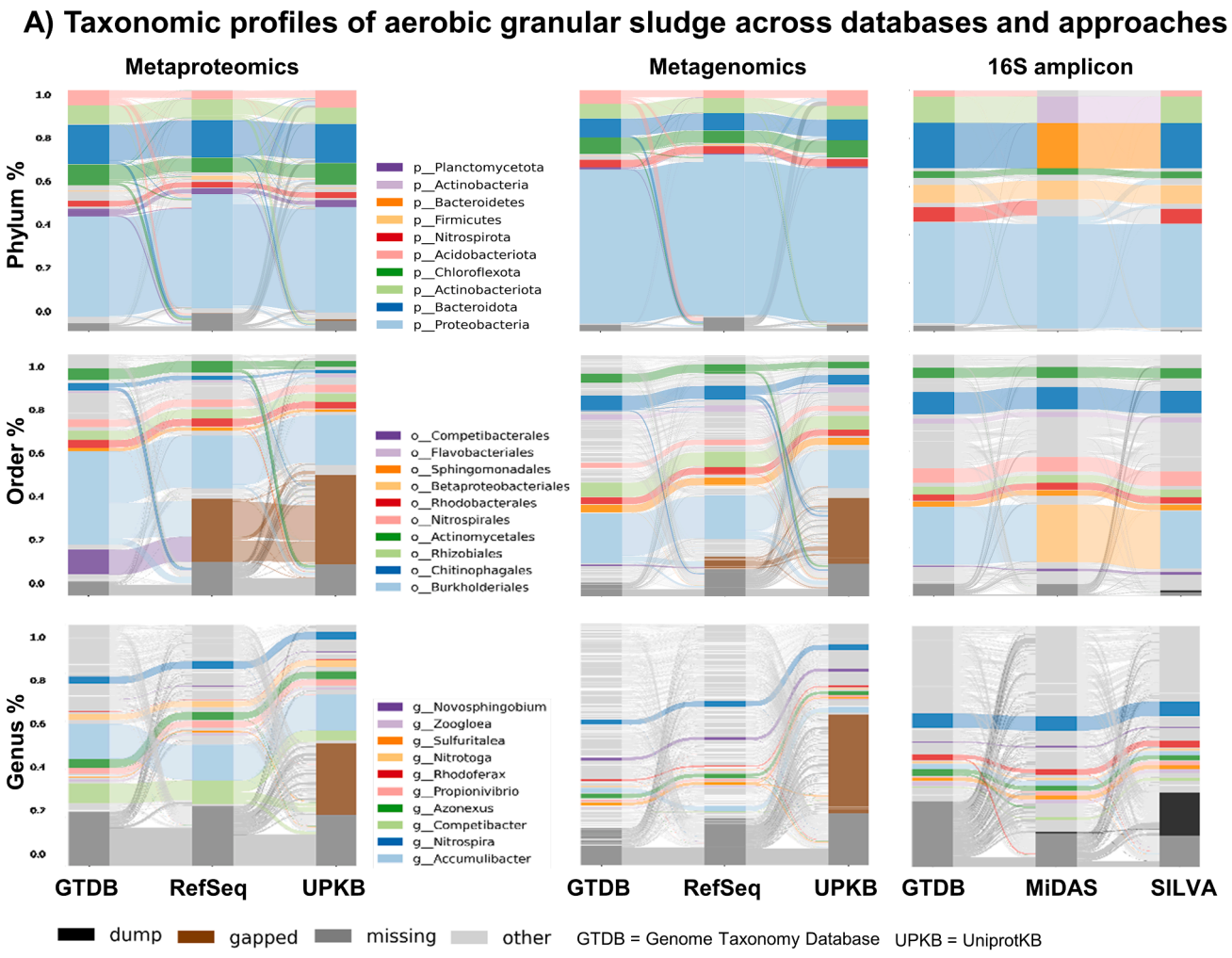


Fig. 2. (A) The Sankey flow diagrams show the impact of different reference sequence databases on taxonomic profiles obtained by metaproteomics, whole metagenome sequencing, and 16S rRNA amplicon sequencing (from the left to right), using a range of different reference sequence databases for taxonomic classification. The Sankey flow diagrams and bar graphs were constructed by combining the annotations obtained from all wastewater treatment plants 1–3. Extended Sankey flow diagrams for metaproteomics detailing all main taxonomic ranks are shown in SI Fig. 11. Fig. 2B) The box plots show the genus fraction variation of the top 10 taxonomies (across all the different reference sequence databases) for metaproteomics, metagenomics and 16S rRNA amplicon sequencing (from left to right). The taxonomic abundances shown in the figures was determined by summing the total number of peptide-to-spectrum matches for metaproteomics, the 'summed average depths of sequencing' for metagenomics, and by using the total ASV counts for 16S rRNA amplicon sequencing.

which values were summed for the compositional analysis.

2.8. Filtering of GTDB for species that contain full length 16S rRNA sequences

The 'normalised' GTDB protein reference sequence database was constructed from organisms which are represented in the 'GTDB ssu reps' (small-subunit ribosomal RNA database) and that contained 'full length' 16S rRNA sequences (sequences with >1200 base pairs were considered as 'full length'). The database normalization procedure is further outlined in SI-doc Sections 2 and 3. The developed python codes for formating the Genome Taxonomy Database for the use with Diamond and QIIME are available at: https://github.com/hbckleikamp/GTDB2DIAMOND | https://github.com/hbckleikamp/GTDB2QIIME.

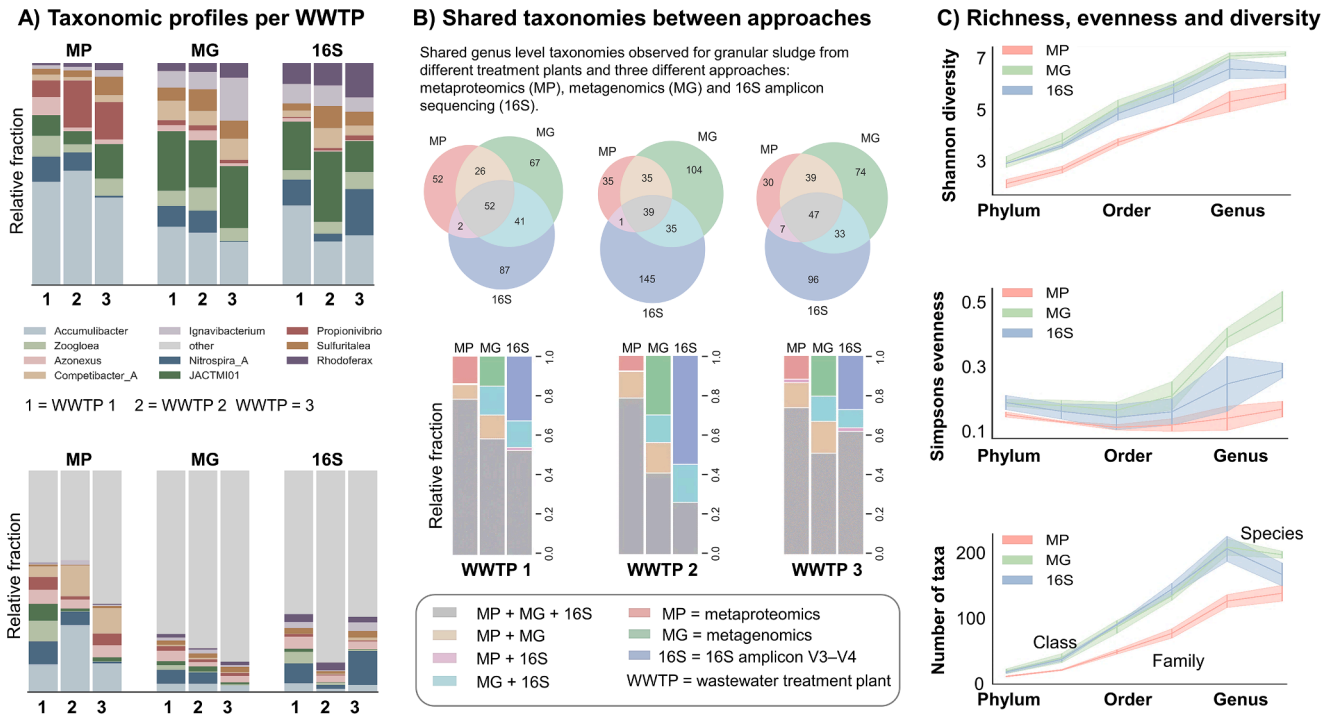


Fig. 3. (A) The bar graphs show the top 10 most abundant genera (when using GTDB for taxonomic classification) obtained by metaproteomics (MP), whole metagenome sequencing (MG), and 16S rRNA amplicon sequencing (16S). The lower bar graphs show the same most abundant genera including other genera grouped as 'other'. The Graphs (B) show shared genera between metaproteomics and the DNA-based approaches, represented as the number of shared taxa (upper Venn diagrams) or as a fraction of the total shared abundance (lower bar graphs). The graphs consider taxa that were observed by at least two techniques and that were present at >3 % abundance, or expressed central nutrient-removing genes. The fraction of genera that were uniformly observed by all three approaches was small compared to the total number of identified taxonomies (grey sections in the Venn diagrams). However, based on metaproteomics, those microbes cover the majority of the protein biomass (grey bars in the bar graphs, labeled with 'MP'). Graphs C) visualizes the microbial diversity indices, including (i) 'Richness,' (ii) 'Simpson's Evenness,' and (iii) 'Shannon diversity,' for the different omics approaches. The data of the individual wastewater treatment plants were averaged. All taxonomic profiles in Fig. A, B, and C were obtained using the corrected GTDB (which contains only taxonomies with 'full-length' 16S reps) for taxonomic classification.

2.9. Taxonomic classification

Differences in the taxonomic composition due to the use of different reference sequence databases were visualized by Sankey flow diagrams. Nomenclature differences between GTDB and NCBI, UniprotKB, SILVA or MiDAS were eliminated using the auxiliary conversion tables obtained from https://data.gtdb.ecogenomic.org/releases/latest/. For example, taxonomic names that matched at least 3 out of 4 times, were changed to the reported name in GTDB. Furthermore, 'Candidatus' prefixes and GTDB unique suffixes such as 'Firmicutes_A', 'Firmicutes_B', were removed. Gaps in the taxonomic lineage annotations were 'bridged' using the name of the closest higher taxonomic rank with a name. The taxonomic abundance in the graphs was calculated from the total ASV counts for 16S amplicon sequencing, the depth of sequencing coverage of contigs for metagenomics, and the total number of peptide-to-spectrum matches for metaproteomics. The employed conversion tables (NCBI and SILVA to GTDB, and vice versa) can be found in the supplementary information (SI-Excel-1–3). However, the stringency of the original CAT algorithm may result in a lower number of genus-level annotations. The algorithm and the parameters were therefore adjusted to improve annotations of dominant taxa, while adhering to 16S amplicon sequencing experiments (SI-doc chapter 1, SI Figs. 1, 2). The taxonomic profiles of the major genera obtained by the original CAT approach and the enhanced CAT approach are shown SI Figs. 3, 4. Interactive Krona charts for the different wastewater treatment plants and the different omics approaches are available via: https://pabstm.github.io/Comparative_metaproteomics_kronas/.

2.10. Shared biomass, diversity, richness and evenness

The shared biomass (at the genus level) was selected from genera that were observed by at least two techniques (=non-unique taxa), and which taxa further were present at >3 % abundance (compared to total abundance of the non-unique taxa within one technique). Taxa which were found to express nutrient-removing genes were included into the evaluation regardless their abundance. Diversity, richness, evenness and shared biomass were determined after uniformly applying an abundance cut-off of 0.1 %. Richness was defined as the number of unique taxa. Simpson's evenness and Shannon's diversity were calculated using the Python functions 'skbio.diversity.alpha.simpson_e(X)' and 'skbio.diversity.alpha_diversity('shannon',X)' which are part of the skbio Python package http://scikit-bio.org. Determination of the abundance of taxa was based on total ASV counts for 16S amplicon, summed depth of sequencing for metagenomics, and the total number of peptide-to-spectrum matches for metaproteomics. Principal coordinate analysis (PCoA) of the Bray–curtis dissimilarity matrix to demonstrate variation in obtained community composition between OMICs approaches and wastewater microbiomes, was performed in python using the function skbio.diversity.beta_diversity, and visualised with a scatter plot. The clustering was performed at different taxonomic levels using the Omics data classified with the filtered GTDB, for taxonomies above a threshold of 0.1 % abundance.

2.11. Functional classification, abundance differences and COG term enrichment analysis

The total abundance was renormalized to a subset of non-unique taxa

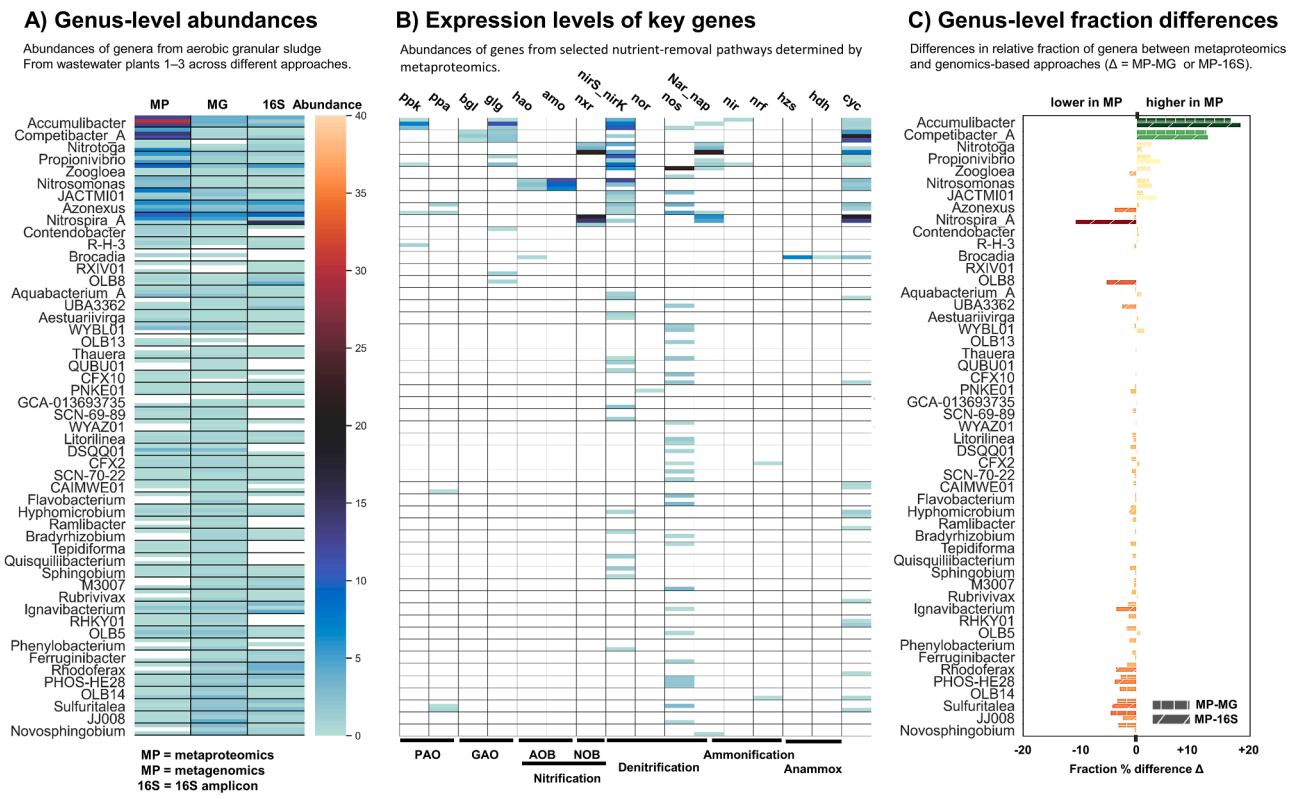


Fig. 4. (A) The heat map shows (key) genera that are present in the aerobic granular sludge microbiome, and which are potentially involved in the central nutrient-removal processes that take place during the wastewater treatment. The taxonomic abundances observed in metaproteomics (MP), metagenomics (MG) and 16S rRNA amplicon sequencing (16S) are shown in separate columns (from left to the right). The abundances observed in the microbiomes obtained from the different treatment plants are shown as individual bars within one cell (top bar = plant 1, middle bar = plant 2 and lower bar = plant 3). Generally, the most dominant genera observed in metaproteomics are *Ca.* Accumulibacter followed by *Ca.* Competibacter. In metagenomics, the most abundant genera are *Nitrospira, Ca.* Accumulibacter and *Azonexus*. Nevertheless, for the genomic approaches, taxonomies were generally found more evenly distributed. (B) The heat map details expression levels of genes from selected nutrient-removal pathways as observed by metaproteomics. The genes are named on the top of the heat map (PPK = polyphosphate kinase, PPA = pyrophosphatase, bglX = beta-glucosidase-like, glg = glycogenin glucosyltransferase, hao = hydroxylamine oxidoreductase, amo = ammonia monooxygenase, nxr = nitrite oxidoreductase, nirK = copper-containing nitrite reductase (EC 2.4.1.186), nirS = cytochrome cd1-containing nitrite reductase, nor = nitric oxide reductase, nos = nitric oxide synthase, nar = respiratory nitrate reductase, nap = periplasmic nitrate reductase, nir = nitrite reductase genes (converting nitrite to nitric oxide), nrf = nitrite reductase (which converts nitrite to ammonium), hzs = hydrazine synthase, hdh = hydrazine dehydrogenase and cyc = cytochrome). The corresponding pathways or organisms are indicated below the heat map (PAO = phosphate accumulating organism, GAO = glycogen accumulating organism, AOB = ammonia-oxidizing bacteria, NOB = nitrite oxidizing bacteria) (C) The graph depicts the relative percentage differences of genera between metaproteomics and metagenomics, represented by bars with a vertical pattern, or between metaproteomics and 16S amplicon sequencing, represented by bars with a diagonal pattern. These differences were calculated by subtracting the normalized genus fraction (%) observed in metagenomics or 16S amplicon sequencing from the normalized genus fraction observed in metaproteomics (Δ = MP-MG or MP-16S). A positive difference (+Δ) indicates a higher relative fraction coverage of the respective genus in metaproteomics experiments, compared to genomics-based approaches, while a negative difference (-Δ) suggests the opposite. The same data are shown as log2 fold fraction differences in SI Fig. 14. The graphs were generated using the normalized Genome Taxonomy Database (GTDB), which only includes taxonomies that are also represented by "full-length" 16S reps.

that showed an abundance of >3 %, or that contained nutrient-removal genes. The between technique absolute abundance difference $(x-y)$ and percent abundance difference $(x-y)/(((x+y))/2)$ was then determined for every genus. Metabolic annotation with KEGG orthologies was performed using BlastKOALA (Kanehisa et al., 2016). Moreover, WEBMGA was used to annotate Clusters of Orthologous Groups (COGs) and protein families (PFAMs and the complementary TIGRFAM terms) (Wu et al., 2011). DIAMOND v2.11 was used to annotate ORFs with UniprotKB genes (Buchfink et al., 2015). The functional analysis and classification was performed by integrating KEGG, COG, PFAM, TIGRFAM and UniprotKB genes (for NXR). Two manually-curated sub-classifications were added to the COG system; 'nitrogen metabolism' (based on KEGG pathways) and 'porin' that includes beta-barrel proteins. Between method COG term enrichment was determined by comparing PSMs from metaproteomic experiments to read counts ('summed sequencing depth') from metagenomics experiments.

3. Results

3.1. A comparative metaomic study on the aerobic granular sludge microbiome

Large-scale omics approaches, such as metaproteomics, metagenomics or 16S rRNA amplicon sequencing, have been rapidly advancing over the past decade. Therefore, efforts have been made in comparing and standardizing procedures. For example, this resulted in the CAMI study for metagenomics (Sczyrba et al., 2017) and in the CAMPI study for metaproteomics (Van Den Bossche et al., 2021). Microbiome studies that integrate different types of approaches are increasingly employed, which also asks for more studies that systematically investigate the complementary character of metaproteomics and DNA-based approaches are urgently needed (Herold et al., 2020; Kleiner et al., 2017; Narayanasamy et al., 2015).

Therefore, we performed a systematic metaomic study on aerobic granular sludge from 3 different wastewater treatment plants. Thereby,

we performed whole metagenome sequencing, metaproteomics and 16S rRNA amplicon sequencing on granules with a uniform size of 2 mm. However, among the sources that significantly limit the comparability between studies and different omics approaches is the existence of different reference sequence databases. Content divergences as well as inaccurate taxonomies and nomenclatures can profoundly impact the taxonomic representation as well as comparability between studies and techniques.

Therefore, we also performed a comparison of the taxonomic profiles and metabolic routes obtained from different reference sequence databases. Nevertheless, a more broadly applicable database with an accurate taxonomy not only allows to more accurately capture the microbial diversity, but it also improves the integration of results from different omics approaches (Godfray, 2002; Parks et al., 2018). Therefore, McDonald et al., established a reference tree that unifies genomic and 16S rRNA databases into a consistent resource (McDonald et al., 2023). Furthermore, the genome taxonomy database (GTDB) uses a set of conserved proteins to normalize taxonomic ranks based on relative evolutionary divergence. This provides an objective, phylogenetically consistent classification of prokaryotes (Chaumeil et al., 2020; Parks et al., 2020, Parks et al., 2022, 2018). Advantageously, GTDB can also be employed to classify the 16S rRNA amplicon sequencing data because it contains small subunit ribosomal RNA sequences (ssu rRNA). Unfortunately, approx. 15 % of the representative taxa in GTDB contain 16S sequences that are shorter than 1200 base pairs and approximately 30 % completely lack corresponding 16S sequences (SI-doc, chapter 2, SI Figs. 5–7).

In order to provide a more broadly applicable database for taxonomic classification of metagenomics, metaproteomics and 16S rRNA amplicon sequencing data, we established a 'filtered' GTDB (SI-doc chapter 2), which contained only organisms with full length 16S rRNA sequences. The Genome Taxonomy Database Toolkit (GTDB-Tk) allows to efficiently classify bacterial and archaeal draft genome assemblies (Chaumeil et al., 2020; Lin et al., 2021; Sedlar et al., 2017). However, in metagenomics, clustering and binning of contigs into genomes commonly results in large unbinned fractions (Chen et al., 2020; Olson et al., 2019). This can significantly bias the taxonomic representation towards the more abundant organisms in a community. In order to provide a more comprehensive sequence database for metaproteomics, we performed the taxonomic classification at the contigs-level. A consensus lineage for each contig was determined using a modified version of the contig annotation tool (CAT) (von Meijenfeldt et al., 2019). The in house developed Python codes for formating the GTDB sequences for the use with DIAMOND and for determining the contig lineages are publicly available (see methods section for Github repository link).

Albeit the filtering for species with full length 16S sequences reduced the number of organisms in the resulting database, the corrected database showed only approx. 5 % less classified reads/PSMs compared to the non-corrected database (SI-doc SI Figures S8–10). Therefore, the reduced database did not significantly impact the taxonomic coverage of the studied aerobic granular sludge microbiome. However, albeit the filtering did not impact the coverage in the present study, it may impact the coverage in other studies, depending on composition of the

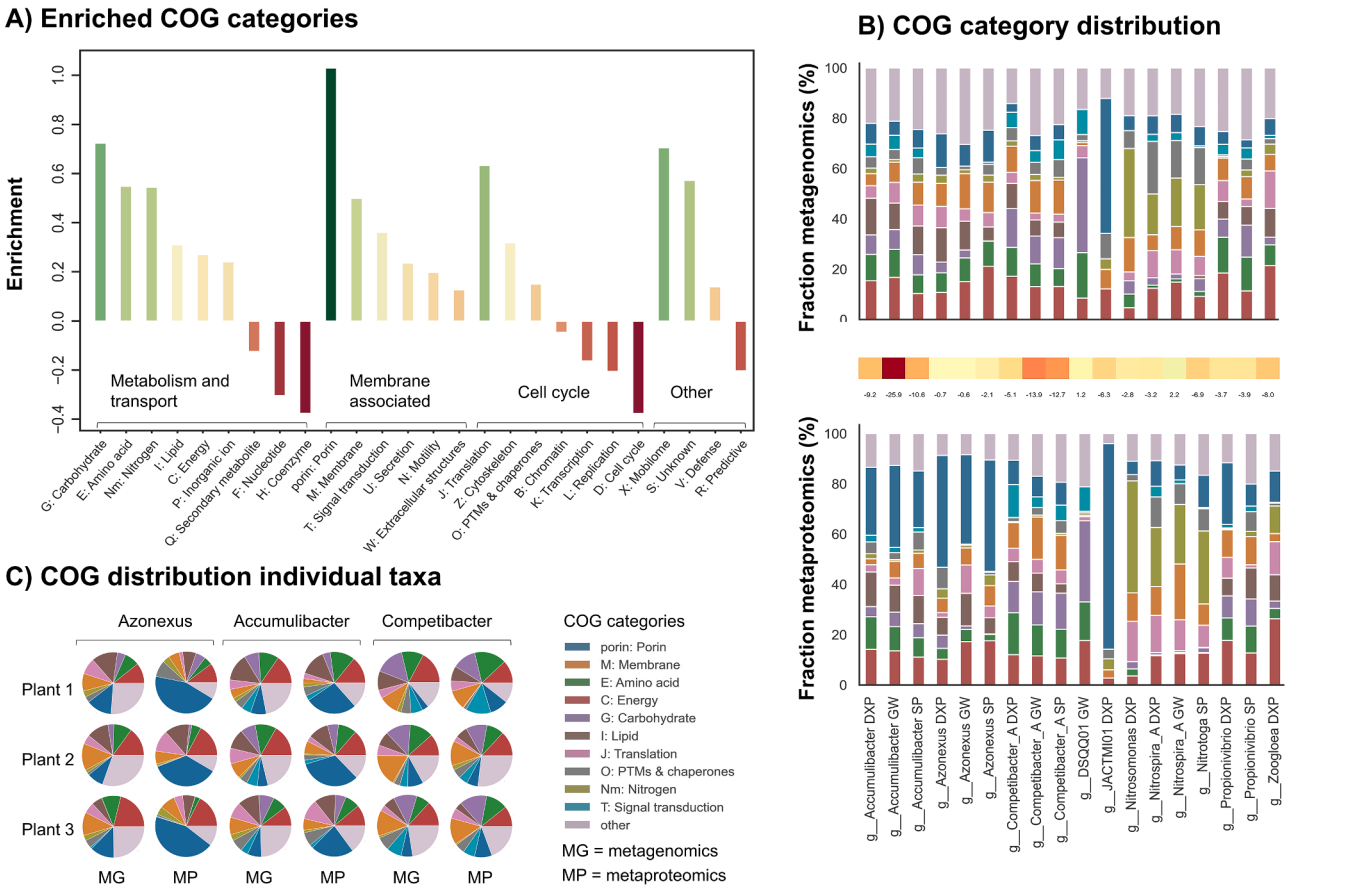


Fig. 5. (A) The bar graph shows the COG term enrichment analysis of the metaproteomics data. (B) The graphs compare the COG category distributions of abundant organisms between metagenomics (upper graph) and metaproteomics (lower graph). (C) The pie charts visualize the proportions of COG categories for selected organisms between metaproteomics (MP) and metagenomics (MG). The metagenomics data are represented as the sum of sequencing depths, while the metaproteomics data are represented by the number of peptide-to-spectrum matches. Graphs A and B display average values from data obtained from the three wastewater treatment plants' microbiomes, while Graph C shows data obtained from the individual treatment plants.

microbial community. Furthermore, 16S-based classification follows a principle which is also not comparable to the whole metagenome, contig-based classification (e.g., Bayesian classifiers are used compared to assembled reads and sequence alignment).

3.2. The impact of reference sequence database divergences on the obtained taxonomic profiles

First, we compared the taxonomic profiles after classifying with different reference sequence databases. For metaproteomics and metagenomics we employed GTDB, RefSeqNR and UniprotKB, and for 16S rRNA we employed GTDB, MiDAS and SILVA. Because the different databases uses individual nomenclatures, the taxonomic names were mapped to GTDB taxonomy (see methods section and SI-DOC). SILVA is transitioning from the NCBI to the GTDB-based taxonomy, which therefore contains lineages from both in addition to unspecific dump taxa. MiDAS is specific to wastewater microbes and uses the AutoTax system (Dueholm 2020), which also contains many MiDAS-exclusive organisms which taxonomies were assigned based on the 16S rRNA genes.

When investigating the obtained taxonomic profiles for metaproteomics and metagenomics, GTDB and RefSeqNR provided the overall highest taxonomic coverage (Fig. 2A). Nevertheless, all three databases provided a comparable relative abundance profile of the main taxonomies. However, there was a decreased level of Competibacter when using UniprotKB. Furthermore, for every experiment there was a substantial fraction of sequences which did not obtain any taxonomic classification, and which therefore evades further interpretation. For metaproteomics using the GTDB this accounted for less than 5 % at the phylum level, and approx. 25 % at the genus level. Moreover, the NCBI-based taxonomies (UniprotKB and RefSeqNR) are not rank-normalized and hence lack certain taxonomic ranks. For example, both Accumulibacter and Competibacter are considered as Candidatus taxa without a family or order name (Oren 2021, “gapped” entries). Moreover, the majority of TrEMBL sequences contain non-curated "dump taxa" of indeterminate taxonomic origin, such as unclassified prokaryotic taxa. Both accounted for a substantial fraction of the taxonomic profiles obtained by metaproteomics and metagenomics.

For the 16S rRNA sequencing all three reference sequence databases, GTDB, MIDAS and SILVA provided nearly the same number of genera. A clear difference however was seen for the genus Tetrasphaera, a major phosphate accumulating genus. This microbe was abundant when using the SILVA and MiDAS database, but it was only poorly annotated when using GTDB. Sequence alignment of the Tetrasphaera ASVs to GTDB demonstrated that a range of other genera were annotated instead of Tetrasphaera (SI-doc, SI Fig. 12, and SI-EXCEL-4). Instead, the Tetrasphaera sequences provided only unspecific annotations at the family-level (Dermatophilaceae). This miss-annotation of Tetrasphaera has been also reported previously (Nouioui et al., 2018; Otieno et al., 2022; Singleton et al., 2022). This demonstrates the limitations when only using V3-V4 16S primers for taxonomic classification. The same was observed for the metaproteomics and metagenomics taxonomic profiles, where Tetrasphaera annotations were nearly absent.

Most interestingly, a comparable set of most abundant taxonomies was observed in metaproteomics and the DNA-based approaches (Fig. 2B). However, the relative fraction of these taxonomies within each omics approach was significantly different. The top 10 taxonomies accounted for nearly 50 % of the sequences in metaproteomics, but only for some 10–15 % in the DNA-based approaches. Furthermore, Competibacter showed a considerably large variation in metaproteomics, where on the other hand *Ca.* Competibacter, Azonexus, Rhodobacer and Rhodoferax showed large variations in metagenomics.

3.3. Taxonomic profiles obtained by different omics approaches

When comparing the taxonomic profiles obtained by the different omics approaches we observe differences already at the phylum level (Fig. 2A). For example, proteobacteria make the most abundant fraction in all the techniques, but their relative fraction is significantly higher in the whole metagenome sequencing data compared to the 16S rRNA data. Differences are even more pronounced at the genus level. For example, *Ca.* Competibacter, *Ca.* Accumulibacter and *Ca.* Nitrotoga are very abundant in the metaproteomics data, but are much less prominent in DNA-based data. On the other hand, the relative fraction of *Nitrospira, Tetrasphaera* and *Rhodoferax* is largest in the 16S rRNA sequencing data. Furthermore, the V3-V4 primers used in this study could not detect *Brocadia* and *Chloroflexota*, which are associated with sludge bulking (Jiang et al., 2021; Speirs et al., 2019).

Nevertheless, regardless of these differences, the relative abundance profiles of the top 10 taxonomies could be considered surprisingly comparable between the different approaches. Also, the relative profile of these taxonomies was comparable between the 3 wastewater treatment plants (Fig. 3A, upper graph). Nonetheless, the fraction of sequences that belonged to the top taxonomies (for GTDB) was only in metaproteomics close to 50 %, but was significantly lower for the DNA-based data (Fig. 3B, lower graph). A large number of low abundant taxonomic identification was also apparent by the moderate number of taxonomies that were consistently identified by all 3 techniques (Fig. 3B, Venn diagrams). On the other hand, the ‘total abundance fraction’ which the shared genera covered was comparatively large (Fig. 3B, grey bars, lower bar graphs). For example, the genera that were observed by all three approaches accounted in metaproteomics for approximately 80 % of the total protein abundance (or biomass).

On the other hand, the fraction that the shared genera covered in 16S rRNA sequencing was significantly lower (approx. 30–60 %). Furthermore, both metagenomics and 16S amplicon sequencing generally showed a larger taxonomic diversity, richness and evenness compared to metaproteomics (Fig. 3C). 16S amplicon sequencing, for example, identified the largest number of taxonomies at the genus level (approx. 200). This was not unexpected, because the DNA-based approaches utilize amplification steps, and also amplify free genetic material, dead and dormant microbial cells. On the other hand, albeit metaproteomics appears to have a lower sensitivity and therefore identified the lowest number of genera, the shared taxonomies (considering the above mentioned thresholds) accounted for a large fraction of the measured protein biomass.

Tables with the obtained abundances for the individual omics approaches and reference sequence databases can be found in the supplementary information (SI-EXCEL-5–7). Moreover interactive Krona charts for all three approaches classified by GTDB (and individual wastewater treatment plants) are available via GitHub: https://pabstm.github.io/Comparative_metaproteomics_kronas/ and the supplementary information as excel macro-enabled workbooks (SI-EXCEL-8–16). The impact of GTDB database normalization on taxonomic profiles for the different omics approaches is shown in SI Figs. 8–10. Extended taxonomic profiles obtained for metaproteomics for all ranks and additional reference sequence database (UniRef100, UniRef90, UniRef50 and Swiss-Prot) are shown in SI Fig. 11.

3.4. Expressed nutrient removal pathways across different wastewater treatment plants

Two key processes of nutrient removal in wastewater treatment are the elimination of nitrogen and phosphorous. To assess genera involved in the conversion of these two core processes we integrated the functional annotations obtained from KEGG, COG terms, PFAM, TIGRFAM domains and UniprotKB genes (Fig. 4). Interestingly, the functional genes covering the nitrogen processes are currently fragmented across different databases. For example, *Nxr* annotation was annotated via UniprotKB, *Nap* by KEGG, and *Nar* using COG terms. Polyphosphate-accumulating organisms (PAO) remove phosphate from the wastewater by producing polyphosphate with the genes *ppk* (polyphosphate

kinase) and *ppa* (pyrophosphatase).

Glycogen-accumulating organisms (GAO)–that compete with PAOs for short-chain fatty acids–synthesize glycogen using *glg* (glycogenin glucosyltransferase; EC 2.4.1.186) and likely therefore show also high expression of *bglX* (beta-glucosidase like enzymes). Nitrogen removal is achieved via subsequent nitrification and denitrification steps that is performed by *hao* (hydroxylamine oxidoreductase) and *amo* (ammonia monooxygenase) genes of ammonia-oxidizing bacteria (AOB) and *nxr* (nitrite oxidoreductase) of nitrite-oxidizing bacteria (NOB). Denitrification (DN) is encoded by the gene clusters *nar* (respiratory nitrate reductase) and *nap* (periplasmic nitrate reductase) to reduce nitrate and *nirK* (copper-containing nitrite reductase) and nirS (cytochrome cd1-containing nitrite reductase) to reduce nitrite, while the genes *nor* (nitric oxide reductase), *nrf* (nitrite reductase) turnover nitric oxide, and ultimately, *nos* (nitric oxide synthase) converts nitrous oxide to dinitrogen gas. Cyc (cytochrome C) is implicated in either the activity of *nor* or *nrf*. Interestingly, *nor* proteins were only detected at low levels, which supposedly is a consequence of membrane association or of poor database annotation accuracy. Furthermore, *hzs* (hydrazine synthase), *hdh* (hydrazine dehydrogenase) as well as *hao* (hydroxylamine oxidoreductase) could be detected in one plant, which are part of the anammox process such as found in *Ca.* Brocadia. Interestingly, several of the key nutrient-removing genera appeared very low abundant in metagenomics and 16S amplicon sequencing data, which was in contrast to the metaproteomics outcomes. These include genera such as *Accumulibacter, Competibacter* and *Propionivibrio* (PAO, GAO and DN, respectively), *Nitrosomonas* (AOB) and *Nitrotoga* (NOB and DN), and *Zoogloea* (DN). Conversely, several other genera, such as *Azonexus* (PAO and DN) and *Nitrospira* (NOB and DN) showed only a minor difference between the orthogonal methods. In addition to *Sulfuritalea* (PAO and DN), other genera were even more prominent in the DNA and rRNA-based approaches. For *Ca.* Accumulibacter, this observation is in agreement with previous studies (Azizan et al., 2020; Barr et al., 2016; Welles et al., 2017), but for *Ca.* Competibacter, however, the observed differences have not been reported before. Moreover, a recent large-scale genomic study showed the widespread presence of genes such as *nosZ* (nitrous-oxide reductase) or *ppk* (polyphosphate kinase), which were detected in a large fraction of the MAGs (Singleton et al., 2021). However, *ppk* for example, could be actually observed by metaproteomics in only a few genera at significant levels. The search terms (used in this study) to extract functional information from the metaproteomics data, as well as a complete table detailing protein taxonomic and functional annotations for all treatment plants can be found in the supplementary information SI-EXCEL documents 17, 18.

3.5. Classification of the observed metaproteome

Proteins make up the bulk of most cells, and thus metaproteomics can be considered as an estimate of the protein biomass composition of microbial communities (Kleikamp et al., 2021; Kleiner et al., 2017). However, we sought to investigate whether the observed abundance differences of individual taxonomies were also affected by the increased detection of specific protein classes. Consequently, we classified the identified proteins by their cluster of orthologous groups (COG) and we included additional groups such as 'nitrogen metabolism' and 'porins'. Furthermore, each COG-group frequency was then compared between proteomics (by considering the number of peptide spectrum matches assigned to this group) and metagenomics (by considering the sequencing coverage assigned to this group). The groups which were found strongly overrepresented in the metaproteomics could be associated with nutrient removal processes (carbohydrate, nitrogen and amino acids), growth (translation) and porins. For example, in Accumulibacter and Azonexus, porins accounted for 30–40 % of peptide matches, while comprising only 5–10 % of the sequencing coverage. This also points to the further presence of outer membrane vesicles in these organisms (Lee et al., 2008). On the other hand, other membrane proteins had an equivalent share in both experiments. However, porins enable the passive transport of a range of molecules, like fatty acids, coenzymes and other small inorganic molecules, therefore are expected to be more abundant in cell membranes. Furthermore, both Nitrospira and Nitrotoga displayed a strong expression of nitrite oxidoreductase (nxr). Competibacter had similar distributions in the metagenomic sequencing coverage and peptide spectrum matches, although it is known to have a large cell volume. Therefore, abundance differences may result from several factors, including over-expression, cell volume, and extraction methods (Albertsen et al., 2015; Pronk et al., 2017).

4. Discussion

Studies that investigate the complementary nature of metaproteomics and the DNA-based approaches for complex environments, such as wastewater treatment plants, are urgently needed. This study presents the first comparative metaomic characterization of the aerobic granular sludge microbiome, sampled from three different wastewater treatment plants. Thereby, we employed (i) metaproteomics, (ii) whole metagenome sequencing, and (iii) 16S rRNA amplicon sequencing to uniform granule material (with a size of 2 mm). Additionally, we investigated the impact of using different reference sequence databases on the taxonomic and functional profiles. Generally, database discrepancies can impede a comparison between studies and different omics approaches. Therefore, we performed our comparison by focusing on the more widely applicable Genome Taxonomy Database (GTDB), which uses a phylogenetically consistent classification of prokaryotes and which contains small subunit ribosomal RNA sequences (ssu rRNA). Although the major taxonomies were consistently identified by all omics approaches, the relative fraction of these taxonomies differed significantly.

For example, the relative fraction of *Ca.* Competibacter, *Ca.* Accumulibacter and *Ca.* Nitrotoga was very high in the metaproteomics data compared to the DNA based approaches. Recent studies already discussed the underrepresentation of *Ca.* Accumulibacter in DNA-based experiments (Azizan et al., 2020; Barr et al., 2016; Kleikamp et al., 2021; Stokholm-Bjerregaard et al., 2017; Welles et al., 2017). Furthermore, DNA-based studies are often challenged with the functional prediction and resolution of strain-level divergences. On the other hand the relative fraction of *Nitrospira, Tetrasphaera* and *Rhodoferax* was high in the DNA based approaches. Tetrasphaera was exclusively detected in 16S rRNA sequencing when using the more specific databases such as SILVA and MiDAS. This bias has been also described by others previously (Nouioui et al., 2018; Otieno et al., 2022; Singleton et al., 2022).

Furthermore, the main taxonomies covered around 50 % of all the sequences in metaproteomics, but only some 10–15 % in the DNA-based approaches. Metaproteomics has been described as a promising approach to estimate the protein biomass distribution (and metabolic capacity) of microbes in communities. Nevertheless, differences in expressed classes of proteins and the commonly employed shotgun experiments may further bias towards the more abundant taxonomies. This was also observed in our study, where metaproteomics detected the lowest number of taxonomies compared to the DNA-based approaches. Yet, compared to DNA-based experiments, metaproteomics uniquely provides insights into the expressed metabolic pathways and enzymes.

Nevertheless, regardless of these differences, the relative abundance profiles of the top taxonomies were surprisingly conserved across the three different wastewater treatment plants. Although the current study aimed to remove biases introduced by the different reference sequence databases, other biases such as discrepancies in DNA and protein extraction procedures, as well as fundamental differences in the sequencing approaches or bioinformatic data processing pipelines were not evaluated in this study. The latter, however, were investigated in more recent lab comparison studies only recently (Sczyrba et al., 2017; Van Den Bossche et al., 2021). Furthermore, this study employed a contig-based taxonomic reference database to increase the

metaproteomic coverage. This focused our study on genus-level resolution, and functional variation of species from the same genus or functional guild were not resolved (Peces et al., 2022). Studies with a focus on functional processes should therefore also consider MAG-based reference sequence databases. Finally, this study aimed to investigate granules with a uniform size of 2 mm. Recent reports, though, have highlighted compositional differences based on the granule size, or flocks respectively (Ali et al., 2019).

5. Conclusions

In this work we provide the first systematic metaproteomic study on the aerobic granular sludge microbiome, which demonstrates the complementary nature of metaproteomics and DNA-based approaches. Our study moreover discusses the importance of generally applicable reference sequence databases, such as GTDB. The application of only one omics approach may thus significantly bias the interpretation of nutrient removal processes. The systematic application of metaproteomics, 16 rRNA sequencing and whole metagenome sequencing as well as the comparison of different reference sequence databases led to the following conclusions:

- While GTDB and RefSeqNR provided the highest taxonomic coverage for metaproteomics and metagenomics, a substantial fraction of sequences did not obtain any taxonomic or functional classifications
- Reference genes for nitrogen processes are currently dispersed among different databases
- Although the number of shared taxonomies was relatively low, the most abundant taxonomies were consistently identified by all omics approaches
- The top 10 taxonomies accounted for approximately 50 % of sequences in metaproteomics, while only 10–15 % in DNA-based approaches
- Metaproteomics showed the lowest diversity, but the consistently identified taxonomies covered approximately 80 % of the measured protein biomass
- The application of single omics approaches, as well as divergences in reference sequence database content and nomenclatures, may profoundly impact the taxonomic and functional interpretation.

The established metaomic data provide a valuable resource for future studies on the metabolic processes in aerobic granular sludge. The omics raw data and Python codes for formating GTDB sequences and the contig-based taxonomic classification are freely accessible through public repositories.

Data availability

The mass spectrometry proteomics raw data have been deposited in the ProteomeXchange consortium database with the dataset identifier PXD030677. Whole metagenome sequencing raw data are available through the NCBI Sequence Read Archive (SRA) under accession numbers SRX13522658– SRX13522660, and the 16S rRNA amplicon sequencing data under the accession numbers SRX21486087–SRX21486101. The BioProject accession number is PRJNA792132. The developed python codes for formating GTDB for the use with Diamond and QIIME are available via https://github.com/hbckleikamp/GTDB2DIAMOND and https://github.com/hbckleikamp/GTDB2QIIME.

Contributions

HK, ML, and MP designed the research, HK, PS and RW performed the sample collection; HK, PS, BA, RW and MP conducted preparation of samples, protein and DNA extraction and sequencing; HK, DG, RV, RZ and MP performed bioinformatic data processing of omics data; HK, MPr and ML analyzed genes involved in nutrient removal; HK, YL, MP, MPr and ML performed (and interpreted) protein enrichment analysis data; HK and MP wrote the manuscript; all authors reviewed the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the data link in the article and SI documents.

Acknowledgments

The authors acknowledge Carol de Ram for support with sample preparation, Leanne van Benthem for collection of the granule materials and Claudia Tugui and all other colleagues from the department of Biotechnology for valuable discussions. The authors acknowledge the SIAM (Soehngen Institute of Anaerobic Microbiology) consortium for funding.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2023.120700.

References

Abbott, S.L., Janda, J.M., 2006. The genus Edwardsiella. Prokaryotes 6, 72–89.
Adav, S.S., Lee, D.J., Lai, J.Y., 2009. Proteolytic activity in stored aerobic granular sludge and structural integrity. Bioresour. Technol. 100 (1), 68–73.
Albertsen, M., Karst, S.M., Ziegler, A.S., Kirkegaard, R.H., Nielsen, P.H., 2015. Back to basics–the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PLoS One 10 (7), e0132783.
Ali, M., Wang, Z., Salam, K.W., Hari, A.R., Pronk, M., van Loosdrecht, M.C., Saikaly, P.E., 2019. Importance of species sorting and immigration on the bacterial assembly of different-sized aggregates in a full-scale aerobic granular sludge plant. Environ. Sci. Technol. 53 (14), 8291–8301.
Angenent, L.T., Karim, K., Al-Dahhan, M.H., Wrenn, B.A., Domíguez-Espinosa, R., 2004. Production of bioenergy and biochemicals from industrial and agricultural wastewater. Trends Biotechnol. 22 (9), 477–485.
Azizan, A., Kaschani, F., Barinas, H., Blaskowski, S., Kaiser, M., Denecke, M., 2020. Using proteomics for an insight into the performance of activated sludge in a lab-scale WWTP. Int. Biodeterior. Biodegrad. 149, 104934.
Balcom, I.N., Driscoll, H., Vincent, J., Leduc, M., 2016. Metagenomic analysis of an ecological wastewater treatment plant's microbial communities and their potential to metabolize pharmaceuticals. F1000Res. 5.
Balvočiūtė, M., Huson, D.H., 2017. SILVA, RDP, Greengenes, NCBI and OTT—How do these taxonomies compare? Bmc Genom. 18 (2), 1–8 [Electronic Resource].
Barr, J.J., Dutilh, B.E., Skennerton, C.T., Fukushima, T., Hastie, M.L., Gorman, J.J., Tyson, G.W., Bond, P.L., 2016. Metagenomic and metaproteomic analyses of accumulibacter phosphatis-enriched floccular and granular biofilm. Environ. Microbiol. 18 (1), 273–287.
Bashiardes, S., Zilberman-Schapira, G., Elinav, E., 2016. Use of Metatranscriptomics in Microbiome Research. Bioinformatics and biology insights 10, BBI, S34610.
Blakeley-Ruiz, J.A., Erickson, A.R., Cantarel, B.L., Xiong, W., Adams, R., Jansson, J.K., Fraser, C.M., Hettich, R.L., 2019. Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes. Microbiome 7 (1), 1–15.
Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., Caporaso, J.G., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6 (1), 1–17.
Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D. A., Caporaso, J.G., 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods 10 (1), 57–59.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15), 2114–2120.
Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37 (8), 852–857.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., Banfield, J.F., 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523 (7559), 208–211.
Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using diamond. Nat. Methods 12 (1), 59–60.
Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7 (5), 335–336.
Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a Toolkit to Classify Genomes With the Genome Taxonomy Database. Oxford University Press.
Chen, L.X., Anantharaman, K., Shaiber, A., Eren, A.M., Banfield, J.F., 2020. Accurate and complete genomes from metagenomes. Genome Res. 30 (3), 315–333.
Cho, I., Blaser, M.J., 2012. The human microbiome: at the interface of health and disease. Nat. Rev. Genet. 13 (4), 260–270.
de Sousa Rollemberg, S.L., de Barros, A.N., Lira, V.N.S.A., Firmino, P.I.M., Dos Santos, A. B., 2019. Comparison of the dynamics, biokinetics and microbial diversity between activated sludge flocs and aerobic granular sludge. Bioresour. Technol. 294, 122106.
den Ridder, M., Daran-Lapujade, P., Pabst, M., 2020. Shot-gun proteomics: why thousands of unidentified signals matter. FEMS Yeast Res. 20 (1) foz088.
Edgar, R.C., 2017. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. PeerJ 5, e3889.
Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27 (16), 2194–2200.
Falkowski, P.G., Fenchel, T., Delong, E.F., 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320 (5879), 1034–1039.
Federhen, S., 2012. The NCBI taxonomy database. Nucleic. Acids. Res. 40 (D1), D136–D143.
Godfray, H.C.J., 2002. Challenges for taxonomy. Nature 417 (6884), 17–19.
Hagen, L.H., Frank, J.A., Zamanzadeh, M., Eijsink, V.G., Pope, P.B., Horn, S.J., Arntzen, M.Ø., 2017. Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. Appl. Environ. Microbiol. 83 (2).
Herold, M., Arbas, S.M., Narayanasamy, S., Sheik, A.R., Kleine-Borgmann, L.A., Lebrun, L.A., Kunath, B.J., Roume, H., Bessarab, I., Williams, R.B., 2020. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. Nat. Commun. 11 (1), 1–14.
Heyer, R., Kohrs, F., Reichl, U., Benndorf, D., 2015. Metaproteomics of complex microbial communities in biogas plants. Microb. Biotechnol. 8 (5), 749–763.
Hugenholtz, P., Skarshewski, A., Parks, D.H., 2016. Genome-based microbial taxonomy coming of age. Cold Spring Harb. Perspect. Biol. 8 (6), a018085.
Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. 11 (1), 1–11.
Integrative, H., Proctor, L.M., Creasy, H.H., Fettweis, J.M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G.A., Snyder, M.P., Strauss III, J.F., 2019. The integrative human microbiome project. Nature 569 (7758), 641–648.
Jansson, J.K., Hofmockel, K.S., 2018. The soil microbiome—From metagenomics to metaphenomics. Curr. Opin. Microbiol. 43, 162–168.
Jiang, C., McIlroy, S.J., Qi, R., Petriglieri, F., Yashiro, E., Kondrotaite, Z., Nielsen, P.H., 2021. Identification of microorganisms responsible for foam formation in mesophilic anaerobic digesters treating surplus activated sludge. Water Res. 191, 116779.
Jouffret, V., Miotello, G., Culotta, K., Ayrault, S., Pible, O., Armengaud, J., 2021. Increasing the power of interpretation for soil metaproteomics data. Microbiome 9 (1), 1–15.
Kanehisa, M., Sato, Y., Morishima, K., 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428 (4), 726–731.
Kehe, J., Kulesa, A., Ortiz, A., Ackerman, C.M., Thakku, S.G., Sellers, D., Kuehn, S., Gore, J., Friedman, J., Blainey, P.C., 2019. Massively parallel screening of synthetic microbial communities. Proc. Natl Acad. Sci. 116 (26), 12804–12809.
Kleikamp, H.B., Pronk, M., Tugui, C., da Silva, Abbas, B., Lin, Y.M., van Loosdrecht, MCM, Pabst, M., 2021. Database-independent de novo metaproteomics of complex microbial communities. Cell Syst. 12 (5), 375–383.
Kleiner, M., 2019. Metaproteomics: much more than measuring gene expression in microbial communities. Msystems 4 (3), e00115–e00119.
Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., Liu, D., Li, C., Strous, M., 2017. Assessing species biomass contributions in microbial communities via metaproteomics. Nat. Commun. 8 (1), 1–14.
Konstantinidis, K.T., Tiedje, J.M., 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187 (18), 6258–6264.
Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 (1), 105–132.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359.
Lawson, C.E., 2021. Retooling microbiome engineering for a sustainable future. Msystems 6 (4) e00925-00921.
Lee, E.Y., Choi, D.S., Kim, K.P., Gho, Y.S., 2008. Proteomics in gram-negative bacterial outer membrane vesicles. Mass Spectrom. Rev. 27 (6), 535–555.
Leventhal, G.E., Boix, C., Kuechler, U., Enke, T.N., Sliwerska, E., Holliger, C., Cordero, O. X., 2018. Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. Nat. Microbiol. 3 (11), 1295–1303.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 (13), 1658–1659.
Li, Z., Wang, Y., Yao, Q., Justice, N.B., Ahn, T.H., Xu, D., Hettich, R.L., Banfield, J.F., Pan, C., 2014. Diverse and divergent protein post-translational modifications in two growth stages of a natural microbial community. Nat. Commun. 5 (1), 1–11.
Liang, Z., Tu, Q., Su, X., Yang, X., Chen, J., Chen, Y., Li, H., Liu, C., He, Q., 2019. Formation, extracellular polymeric substances, and structural stability of aerobic granules enhanced by granular activated carbon. Environ. Sci. Pollut. Res 26 (6), 6123–6132.
Lin, Y., Wang, L., Xu, K., Li, K., Ren, H., 2021. Revealing taxon-specific heavy metal-resistance mechanisms in denitrifying phosphorus removal sludge using genome-centric metaproteomics. Microbiome 9 (1), 1–17.
Lohmann, P., Schäpe, S.S., Haange, S.B., Oliphant, K., Allen-Vercoe, E., Jehmlich, N., Von Bergen, M., 2020. Function is what counts: how microbial community complexity affects species, proteome and pathway coverage in metaproteomics. Expert Rev. Proteom. 17 (2), 163–173.
Louca, S., Doebeli, M., Parfrey, L.W., 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6 (1), 1–12.
Lovley, D.R., 2017. Happy together: microbial communities that hook up to swap electrons. ISME J. 11 (2), 327–336.
Magoč, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27 (21), 2957–2963.
May, D.H., Timmins-Schiffman, E., Mikan, M.P., Harvey, H.R., Borenstein, E., Nunn, B.L., Noble, W.S., 2016. An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. J. Proteome Res. 15 (8), 2697–2705.
McDonald, D., Jiang, Y., Balaban, M., Cantrell, K., Zhu, Q., Gonzalez, A., Morton, J.T., Nicolaou, G., Parks, D.H., Karst, S.M., 2023. Greengenes2 unifies microbial data in a single reference tree. Nat. Biotechnol. 1–4.
McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6 (3), 610–618.
Morrissey, E.M., Mau, R.L., Schwartz, E., Caporaso, J.G., Dijkstra, P., Van Gestel, N., Koch, B.J., Liu, C.M., Hayer, M., McHugh, T.A., 2016. Phylogenetic organization of bacterial activity. ISME J. 10 (9), 2336–2340.
Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., Martens, L., Renard, B. Y., 2018. MPA portable: a stand-alone software package for analyzing metaproteome samples on the go. Anal. Chem. 90 (1), 685–689.
Narayanasamy, S., Muller, E.E., Sheik, A.R., Wilmes, P., 2015. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. Microb. Biotechnol. 8 (3), 363–368.
Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J.P., Woyke, T., Kyrpides, N.C., Pukall, R., Klenk, H.P., Goodfellow, M., Göker, M., 2018. Genome-based taxonomic classification of the phylum Actinobacteria. Front. Microbiol. 9, 2007.
Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27 (5), 824–834.
Okonechnikov, K., Conesa, A., García-Alcalde, F., 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32 (2), 292–294.
Olson, N.D., Treangen, T.J., Hill, C.M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., Pop, M., 2019. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief. Bioinform. 20 (4), 1140–1150.
Orhon, D., Babuna, F.G., Karahan, O., 2009. Industrial Wastewater Treatment By Activated Sludge. IWA Publishing.
Otieno, J., Kowal, P., Mąkinia, J., 2022. The occurrence and role of tetrasphaera in enhanced biological phosphorus removal systems. Water 14 (21), 3428 (Basel).
Pabst, M., Grouzdev, D.S., Lawson, C.E., Kleikamp, H.B., de Ram, C., Louwen, R., Lin, Y. M., Lücker, S., van Loosdrecht, M., Laureni, M., 2021. A general approach to explore prokaryotic protein glycosylation reveals the unique surface layer modulation of an anammox bacterium. ISME J. 1–12.
Panchavinin, S., Tobino, T., Hara-Yamamura, H., Matsuura, N., Honda, R., 2019. Candidates of quorum sensing bacteria in activated sludge associated with N-acyl homoserine lactones. Chemosphere 236, 124292.
Parks, D.H., Chuvochina, M., Chaumeil, P.A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. 38 (9), 1079–1086.
Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50 (D1), D785–D794.
Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36 (10), 996–1004.
Peces, M., Dottorini, G., Nierychlo, M., Andersen, K.S., Dueholm, M.K.D., Nielsen, P.H., 2022. Microbial communities across activated sludge plants show recurring species-level seasonal patterns. ISME Commun. 2 (1), 18.
Pronk, M., De Kreuk, M., De Bruin, B., Kamminga, P., Kleerebezem, R.v., Van Loosdrecht, M., 2015. Full scale performance of the aerobic granular sludge process for sewage treatment. Water Res. 84, 207–217.
Pronk, M., Neu, T.R., Van Loosdrecht, M., Lin, Y., 2017. The acid soluble extracellular polymeric substance of aerobic granular sludge dominated by defluviicoccus sp. Water Res. 122, 148–158.
Püttker, S., Kohrs, F., Benndorf, D., Heyer, R., Rapp, E., Reichl, U., 2015. Metaproteomics of activated sludge from a wastewater treatment plant–A pilot study. Proteomics 15 (20), 3596–3601.
Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N., 2017. Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35 (9), 833–844.
Rabaey, K., Verstraete, W., 2005. Microbial fuel cells: novel biotechnology for energy generation. Trends Biotechnol. 23 (6), 291–298.

Ramos, C., Suárez-Ojeda, M.E., Carrera, J., 2015. Long-term impact of salinity on the performance and microbial population of an aerobic granular reactor treating a high-strength aromatic wastewater. Bioresour. Technol. 198, 844–851.
Ranjan, R., Rani, A., Metwally, A., McGee, H.S., Perkins, D.L., 2016. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. 469 (4), 967–977.
Rousk, J., Bengtson, P., 2014. Microbial regulation of global biogeochemical cycles. Front. Microbiol. 5, 103.
Rubio-Rincón, F., Weissbrodt, D., Lopez-Vazquez, C., Welles, L., Abbas, B., Albertsen, M., Nielsen, P., Van Loosdrecht, M., Brdjanovic, D., 2019. Candidatus Accumulibacter delftensis": a clade IC novel polyphosphate-accumulating organism without denitrifying activity on nitrate. Water Res. 161, 136–151.
Salvato, F., Hettich, R.L., Kleiner, M., 2021. Five key aspects of metaproteomics as a tool to understand functional interactions in host-associated microbiomes. PLoS Pathog. 17 (2), e1009245.
Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., Karsch-Mizrachi, I., 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020, baaa062.
Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., 2017. Critical assessment of metagenome interpretation—A benchmark of metagenomics software. Nat. Methods 14 (11), 1063–1071.
Sedlar, K., Kupkova, K., Provaznik, I., 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput. Struct. Biotechnol. J. 15, 48–55.
Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P., 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15 (2), 121–132.
Singleton, C., Petriglieri, F., Wasmund, K., Nierychlo, M., Kondrotaite, Z., Petersen, J., Peces, M., Dueholm, M., Wagner, M., Nielsen, P., 2022. The novel genus,'Candidatus Phosphoribacter', previously identified as Tetrasphaera, is the dominant polyphosphate accumulating lineage in EBPR wastewater treatment plants worldwide. ISME J. 16 (6), 1605–1616.
Singleton, C.M., Petriglieri, F., Kristensen, J.M., Kirkegaard, R.H., Michaelsen, T.Y., Andersen, M.H., Kondrotaite, Z., Karst, S.M., Dueholm, M.S., Nielsen, P.H., 2021. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat. Commun. 12 (1), 1–13.
Speirs, L.B., Rice, D.T., Petrovski, S., Seviour, R.J., 2019. The phylogeny, biodiversity, and ecology of the Chloroflexi in activated sludge. Front. Microbiol. 10, 2015.
Starke, R., Pylro, V.S., Morais, D.K., 2021. 16S rRNA gene copy number normalization does not provide more reliable conclusions in metataxonomic surveys. Microb. Ecol. 81 (2), 535–539.
Stoddard, S.F., Smith, B.J., Hein, R., Roller, B.R., Schmidt, T.M., 2015. rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 43 (D1), D593–D598.
Stokholm-Bjerregaard, M., McIlroy, S.J., Nierychlo, M., Karst, S.M., Albertsen, M., Nielsen, P.H., 2017. A critical assessment of the microorganisms proposed to be important to enhanced biological phosphorus removal in full-scale wastewater treatment systems. Front. Microbiol. 8, 718.
Świątczak, P., Cydzik-Kwiatkowska, A., 2018. Performance and microbial characteristics of biomass in a full-scale aerobic granular sludge wastewater treatment plant. Environ. Sci. Pollut. Res. 25 (2), 1655–1669.
Szabó, E., Liébana, R., Hermansson, M., Modin, O., Persson, F., Wilén, B.M., 2017. Comparison of the bacterial community composition in the granular and the suspended phase of sequencing batch reactors. AMB Express 7 (1), 1–12.
Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M.F., 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome 4 (1), 1–13.
Tawalbeh, M., Al-Othman, A., Singh, K., Douba, I., Kabakebji, D., Alkasrawi, M., 2020. Microbial desalination cells for water purification and power generation: a critical review. Energy 209, 118493.
Temudo, M.F., Muyzer, G., Kleerebezem, R., van Loosdrecht, M.C., 2008. Diversity of microbial communities in open mixed culture fermentations: impact of the pH and carbon source. Appl. Microbiol. Biotechnol. 80 (6), 1121–1130.
Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I., 2007. The human microbiome project. Nature 449 (7164), 804–810.
Van Den Bossche, T., Kunath, B.J., Schallert, K., Schäpe, S.S., Abraham, P.E., Armengaud, J., Arntzen, M.Ø., Bassignani, A., Benndorf, D., Fuchs, S., Giannone, R. J., Griffin, T.J., Hagen, L.H., Halder, R., Henry, C., Hettich, R.L., Heyer, R., Jagtap, P., Jehmlich, N., Jensen, M., Juste, C., Kleiner, M., Langella, O., Lehmann, T., Leith, E., May, P., Mesuere, B., Miotello, G., Peters, S.L., Pible, O., Queiros, P.T., Reichl, U., Renard, B.Y., Schiebenhoefer, H., Sczyrba, A., Tanca, A., Trappe, K., Trezzi, J.P., Uzzau, S., Verschaffelt, P., von Bergen, M., Wilmes, P., Wolf, M., Martens, L., Muth, T., 2021. Critical assessment of meta proteome investigation (CAMPI): a multi-laboratory comparison of established workflows. Nat. Commun 12 (1), 7305.
van Loosdrecht, M.C., Brdjanovic, D., 2014. Anticipating the next century of wastewater treatment. Science 344 (6191), 1452–1453.
von Meijenfeldt, F.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., Dutilh, B.E., 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol. 20 (1), 1–14.
Weissbrodt, D.G., Neu, T.R., Kuhlicke, U., Rappaz, Y., Holliger, C., 2013. Assessment of bacterial and structural dynamics in aerobic granular biofilms. Front. Microbiol. 4, 175.
Weissbrodt, D.G., Shani, N., Holliger, C., 2014. Linking bacterial population dynamics and nutrient removal in the granular sludge biofilm ecosystem engineered for wastewater treatment. FEMS Microbiol. Ecol. 88 (3), 579–595.
Welles, L., Abbas, B., Sorokin, D.Y., Lopez-Vazquez, C.M., Hooijmans, C.M., van Loosdrecht, M., Brdjanovic, D., 2017. Metabolic response of "Candidatus Accumulibacter Phosphatis" clade II C to changes in influent P/C ratio. Front. Microbiol. 7, 2121.
Welles, L., Tian, W., Saad, S., Abbas, B., Lopez-Vazquez, C., Hooijmans, C., Van Loosdrecht, M., Brdjanovic, D., 2015. Accumulibacter clades Type I and II performing kinetically different glycogen-accumulating organisms metabolisms for anaerobic substrate uptake. Water Res. 83, 354–366.
Wilmes, P., Heintz-Buschart, A., Bond, P.L., 2015. A decade of metaproteomics: where we stand and what the future holds. Proteomics 15 (20), 3409–3417.
Wilmes, P., Wexler, M., Bond, P.L., 2008. Metaproteomics provides functional insight into activated sludge wastewater treatment. PLoS One 3 (3), e1778.
Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., Zhang, Q., Brown, M.R., Li, Z., Van Nostrand, J.D., 2019. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat. Microbiol. 4 (7), 1183–1195.
Wu, S., Zhu, Z., Fu, L., Niu, B., Li, W., 2011. WebMGA: a customizable web server for fast metagenomic sequence analysis. Bmc Genom. 12 (1), 1–9 [Electronic Resource].
Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzéby, J., Amann, R., Rosselló-Móra, R., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12 (9), 635–645.
Zhang, T., Shao, M.F., Ye, L., 2012. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. ISME J. 6 (6), 1137–1147.
Zhou, J., Sun, Q., 2020. Performance and microbial characterization of aerobic granular sludge in a sequencing batch reactor performing simultaneous nitrification, denitrification and phosphorus removal with varying C/N ratios. Bioprocess. Biosyst. Eng. 43 (4), 663–672.
Zorz, J.K., Sharp, C., Kleiner, M., Gordon, P.M., Pon, R.T., Dong, X., Strous, M., 2019. A shared core microbiome in soda lakes separated by large distances. Nat. Commun. 10 (1), 1–10.