Comprehensive ecosystem-specific 16S rRNA gene databases with automated taxonomy assignment (AutoTax) provide species-level resolution in microbial ecology

Authors: Morten Simonsen Dueholm, Kasper Skytte Andersen, Francesca Petriglieri, Simon Jon McIlroy, Marta Nierychlo, Jette Fisher Petersen, Jannie Munk Kristensen, Erika Yashiro, Søren Michael Karst, Mads Albertsen, and Per Halkjær Nielsen*

Affiliation:

Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark.

*Correspondence to: Per Halkjær Nielsen, Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220 Aalborg, Denmark; Phone: +45 9940 8503; Fax: Not available; E-mail: phn@bio.aau.dk

Running title:

Comprehensive ecosystem-specific 16S rRNA gene databases with automated taxonomy assignment

Abstract: High-throughput 16S rRNA gene amplicon sequencing is an indispensable method for studying the diversity and dynamics of microbial communities. However, this method is presently hampered by the lack of high-identity reference sequences for many environmental microbes in the public 16S rRNA gene reference databases, and by the lack of a systematic and comprehensive taxonomic classification for most environmental bacteria. Here we combine high-quality and high-throughput full-length 16S rRNA gene sequencing with a novel sequence identity-based approach for automated taxonomy assignment (AutoTax) to create robust, near-complete 16S rRNA gene databases for complex environmental ecosystems. To demonstrate the benefit of the approach, we created an ecosystem-specific database for wastewater treatment systems and anaerobic digesters. The novel approach allows consistent species-level classification of 16S rRNA amplicons sequence variants and the design of highly specific oligonucleotide probes for fluorescence *in situ* hybridization, which can reveal *in situ* properties of microbes at unprecedented taxonomic resolution.

Introduction

Microbial communities determine the functions of microbial ecosystems in nature and engineered systems. A deep understanding of the communities requires reliable identification of the microbes present, as well as linking their identity with functions. Identification at the lowest taxonomic rank is preferred, as microbial traits vary in their degree of phylogenetic conservation, and many ecologically important traits are conserved only at the family to species rank[1].

Identification of microbes is commonly achieved by high throughput 16S rRNA gene amplicon sequencing, where a segment of the 16S rRNA gene spanning one to three hypervariable regions is amplified by PCR and sequenced. The amplicons are then clustered, based on sequence identity into operational taxonomic units (OTUs) or used to infer exact amplicon sequence variants (ASVs), also commonly known as sub-OTUs (sOTUs) or zero-radius OTUs (zOTUs), with denoising algorithms such as Deblur[2], DADA2[3], and Unoise3[4]. The sequences are finally classified, based on a 16S rRNA gene reference database to assign the most plausible taxonomy for each sequence[5]. ASVs are often preferred over OTUs, because they provide the highest phylogenetic resolution, supporting sub-genus to sub-species level classification, depending on the 16S rRNA gene region amplified and the taxon analyzed[6].

ASVs can be applied as consistent labels for microbial identification independently of a 16S rRNA gene reference database[6]. This approach is used in several large-scale projects, including the Earth Microbiome Project (EMP)[7] and the American Gut project[8], to provide detailed insight into the factors that shape the overall microbial community diversity and dynamics. However, ASVs are not ideal as references for linking microbial identity with functions. Firstly, ASVs do not contain enough evolutionary information to confidently resolve their phylogeny[9,10], which makes it impossible to report and infer how microbial traits are conserved at different phylogenetic scales. Secondly, comparison of ASVs is only possible when they are produced and processed in the same way. This means that, without taxonomic assignment, it is not possible to compare results across studies that have used primer sets targeting different regions of the 16S rRNA gene. It also hampers our ability to exploit the power of new or improved sequencing technologies that can produce longer reads of high quality. Finally, if information about functional properties is available from pure cultures studies or *in situ* based on metagenome assembled genomes (MAGs), this information may be linked to full-length 16S rRNA sequences, but less reliably to ASVs[9,10]. Taxonomic assignment is therefore crucial for cross-study comparisons and the dissemination of microbial knowledge.

Taxonomic assignment to ASVs relies on the classifier (e.g., sintax or RDP classifier) that applies different algorithms to compare each individual ASV to a 16S rRNA gene reference database and proposes the best estimate for the taxonomy. Confident classification at the lowest taxonomic ranks requires high-identity reference sequences (~100% identity) and a comprehensive taxonomy for all references[10]. None of these criteria are met with the commonly applied universal reference databases (Greengenes[11], SILVA[12], and RDP[13]), which lack sequences for many environmental taxa and a comprehensive taxonomy for most uncultivated taxa.

A solution to the aforementioned problems is to create ecosystem-specific reference databases. Some well-studied medium-complexity ecosystems, such as the human gut or oral microbiomes, now have fairly comprehensive reference databases with genus- to species-level resolution, which were obtained from thousands of isolates and MAGs[14–16]. However, this is not yet the case for most environmental ecosystems.

New methods for high throughput full-length 16S rRNA gene sequencing, e.g., synthetic long-read sequencing on the Illumina platform[17,18], but also emerging methods such as PacBio[19] and Nanopore[20] consensus sequencing, now allow generation of millions of high-quality reference sequences from any environmental ecosystem. This can provide high-identity references for many of the uncultured taxa which are currently missing in the large universal reference databases, but does not solve the problem of missing or poor taxonomic assignment for many taxa.

Current strategies for generating and maintaining ecosystem-specific taxonomies involve ecosystem-specific curated versions of universal reference databases, where the taxonomy is manually curated for some process-critical microbes, and placeholder names are provided for the most abundant uncultured genera. Examples are the MiDAS database for microbes in biological wastewater treatment systems[21] and smaller ecosystem-specific databases that only include sequences from the specific ecosystem, with the taxonomy rigorously curated by scientists within the field such as the freshwater-specific FreshTrain database[22], the human intestinal 16S rRNA gene taxonomic database[14], and the human oral microbiome database[16]. However, manual ecosystem-specific curation of the reference databases is subjective and hardly sustainable with the fast growing number of sequences in such databases[23].

Ideally, we want an automated taxonomy assignment which can provide robust, objective taxonomic classifications for all ESVs, based on the most recent microbial taxonomy with introduction of placeholder names for taxa which have not yet received official names. To achieve this, we introduce AutoTax - a simple and efficient strategy to create a comprehensive ecosystem-specific taxonomy covering all taxonomic ranks. AutoTax uses the SILVA taxonomy as a backbone and provides robust placeholder names for unclassified taxa, based on *de novo* clustering of sequences according to statistically supported identity thresholds for each taxonomic rank[9]. Due to the strict computational nature of the taxonomy assignment, we obtain an objective taxonomy, which can easily be updated, based on the most recent taxonomy in other reference databases.

We demonstrate the potential of the method by sequencing almost a million full-length small subunit rRNA gene (fSSU) sequences from Danish bioenergy and biological wastewater treatment systems and use these after error correction to create a new comprehensive ecosystem-specific reference database with 21,039 exact sequence variants (ESVs), which were classified using AutoTax. The value of the new approach was demonstrated by comparing the performance with the large universal reference database commonly applied. The comprehensive set of ESVs also allowed the design of species or sequence variant-specific oligonucleotide probes for fluorescence *in situ* hybridization (FISH). This was exemplified by new probes for one of the most abundant genera in Danish wastewater treatment systems, the *Tetrasphaera*, where it enabled the visual distinction of several species revealing different phenotypes.

Materials and methods:

General molecular methods

Concentration and quality of nucleic acids were determined using a Qubit 3.0 fluorometer (Thermo Fisher Scientific) and an Agilent 2200 Tapestation (Agilent Technologies), respectively. Agencourt RNAClean XP and AMPure XP beads were used as described by the manufacturer, except for the washing steps, where 80% ethanol was used. RiboLock RNase inhibitor (Thermo Fisher Scientific) was added to the purified total RNA to minimise RNA degradation. All commercial kits were used according to the protocols provided by the manufacturer, unless otherwise stated. Oligonucleotides used in this study can be found in **Table S1**.

Samples and nucleic purification

Activated sludge and anaerobic digester biomass were obtained as frozen aliquots (-80°C) from the MiDAS collection[21]. Sample metadata is provided in **Table S2**. Total nucleic acids were purified from 500 µL of sample thawed on ice using the PowerMicrobiome RNA isolation kit (MO BIO Laboratories) with the optional phenol-based lysis or with the RiboPure RNA purification kit for bacteria (Thermo Fisher Scientific). Purification was carried out according to the manufacturers' recommendations, except that cell lysis was performed in a FastPrep-24 instrument for 4x 40 s at 6.0 m/s to increase the yield of nucleic acids from bacteria with tough cell walls[24]. The samples were incubated on ice for 2 min between each bead beating to minimise heating due to friction. DNA-free total RNA was obtained by treating a subsample of the purified nucleic acid with the DNase Max kit (MO BIO Laboratories), followed by clean up using 1.0x RNAClean XP beads with elution into 25 µL nuclease-free water.

Primer-free full-length 16S rRNA library preparation and sequencing

Purified RNA obtained from biomass samples was pooled for each sample source type (activated sludge or anaerobic digester) to give equimolar amounts of the small subunit ribosomal ribonucleic acid (SSU rRNA) determined based on peak area in the TapeStation analysis software A.02.02 (SR1). Full-length SSU sequencing libraries were then prepared as previously described[17]. The SSU_rRNA_RT2 (activated sludge) and SSU_rRNA_RT3 (anaerobic digester biomass) reverse transcription primer and the SSU_rRNA_l adaptor were used for the molecular tagging, and approximately 1,000,000 tagged molecules from each pooled sample were used to create the clonal library. The final library was sequenced on a HiSeq2500 using on-board clustering and rapid run mode with a HiSeq PE Rapid Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2, 265 cycles (Illumina), as previously described[17].

Primer-based full-length 16S rRNA library preparation and sequencing

The purified nucleic acids obtained from the biomass samples were pooled for each sample source type (activated sludge or anaerobic digester) with equal weight of DNA from each sample. Full-length SSU sequencing libraries were then prepared, as previously described[17]. The f16S_rDNA_pcr1_fw1 (activated sludge) or f16S_rDNA_pcr1_fw2 (anaerobic digester biomass) and the f16S_rDNA_pcr1_rv were used for the molecular tagging, and approximately 1,000,000 tagged molecules from each pooled sample were used to create the clonal library. The final library was sequenced on a HiSeq2500 using on-board clustering and rapid run mode with a HiSeq PE Rapid Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2, 265 cycles (Illumina) as previously described[17].

Preparation of full-length 16S rRNA gene exact sequence variants (ESVs)

Raw sequence reads were binned, based on the unique molecular tags, *de novo* assembled into the synthetic long-read rRNA gene sequences using the fSSU-pipeline-DNA_v1.2.sh or fSSU-pipeline-RNA_v1.2.sh scripts script (https://github.com/KasperSkytte/AutoTax)[17]. The assembled 16S rRNA gene sequences were trimmed equivalent to *E. coli* position 8 and 1507 (RNA-based protocol) or 28 and 1491 (DNA-based protocol), as previously described[17]. This ensures that the sequences have equal length and that primer binding sites are removed from the DNA-based sequences. Exact sequence variants (ESVs) were obtained by identifying unique sequences, which were observed at least twice, and discarding shorter ESVs that match exactly with longer ESVs using the ESVpipeline.sh shell script (https://github.com/KasperSkytte/AutoTax). For details see the supplementary results.

Taxonomy assignment to ESVs

A complete taxonomy from kingdom to species was automatically assigned to each ESV using the AutoTax.sh scripts (https://github.com/KasperSkytte/AutoTax). In brief, this script identifies the closest relative of each ESV in the SILVA database, obtains the taxonomy for this sequence, and discards information at taxonomic ranks not supported by the sequence identity, based on the thresholds for taxonomic ranks proposed by Yarza *et al.*[9]. In addition, ESVs are *de novo* clustered using the UCLUST algorithm and the same thresholds. The *de novo* clusters are labelled based on the centroid ESV, and these labels act as a placeholder taxonomy, where there are gaps in the taxonomy obtained from SILVA. For details, see the supplementary results.

Amplicon sequencing and analysis

Bacterial community analysis was performed by amplicon sequencing of the V1-3 variable region as previous described[25] using the 27F (AGAGTTTGATCCTGGCTCAG[26]) and 534R (ATTACCGCGGCTGCTGG[27]) primers and the purified DNA from above. Forward reads were processed using usearch v.11.0.667. Raw fastq files were filtered for phiX sequences using -filter_phix, trimmed to 250 bp using -fastx_truncate -trunclen 250, and quality filtered using -fastq_filter with -fastq_maxee 1.0. The sequences were dereplicated using -fastx_uniques with -sizeout -relabel Uniq. Exact amplicon sequence variants (ASVs) were generated using -unoise3[4]. ASV-tables were created by mapping the raw reads to the ASVs using -otutab with the -zotus and -strand both options. Taxonomy was assigned to ASVs using -sintax with -strand both and -sintax_cutoff 0.8[10].

Data analysis and visualization

Usearch v.10. 0.240 was used for mapping sequences to references with -usearch_global -id 0 -maxrejects 0 -strand plus, unless otherwise stated. Data was imported into R[28] using RStudio IDE[29], analysed, and aggregated using Tidyverse v.1.2.1 (https://www.tidyverse.org/), and visualised using ggplot2[30] v.3.1.0 and Ampvis[31] v.2.4.0.

Data availability

Raw and assembled sequencing data is available at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under the project number PRJEB26558.

Fluorescence in situ hybridization (FISH)

Fresh biomass samples from full-scale activated sludge WWTP were fixed with 96% ethanol and stored in the freezer (-20°C) until needed. FISH was performed as described by Daims et al.[32]. Details about the optimal formamide concentration used for each probe are given in **Table S4**. The EUBmix probe set[33,34] was used to cover all bacteria, and the nonsense NON-EUB probe[35] was applied as negative control for sequence-independent probe binding. Microscopic analysis was performed with either an Axioskop epifluorescence microscope (Carl Zeiss, Germany), equipped with a Leica DFC7000 T CCD camera, or a white light laser confocal microscope (Leica TCS SP8 X) (Leica Microsystems, Wetzlar, Germany).

Phylogenetic analysis and FISH probe design

Phylogenetic analysis of 16S rRNA gene sequences and the design of FISH probes for individual species in the genus *Tetrasphaera* were performed using the ARB software v.6.0.6[36]. A phylogenetic tree was calculated, based on the aligned 722 new ESVs from the genus *Tetrasphaera*, using the PhyML maximum likelihood method and a 1000-replicate bootstrap analysis. Unlabelled helper probes and competitor probes were designed for regions predicted to have low in situ accessibility and for single base mismatched non-target sequences, respectively. Potential probes were validated *in silico* with the MathFISH software for hybridization efficiencies of target and potentially weak non-target matches[37]. All probes were purchased from Sigma-Aldrich (Denmark) or Biomers (Germany), labelled with 6-carboxyfluorescein (6-Fam), indocarbocyanine (Cy3) or indodicarbocyanine (Cy5) fluorochromes. Optimal hybridization conditions for novel

FISH probes were determined, based on formamide dissociation curves, generated after hybridization at different formamide concentrations over a range of 0–70% (v/v) with 5% increments. Relative fluorescence intensities of 50 cells were measured with the ImageJ software (National Institutes of Health, Maryland, USA) and calculated average values were compared for selection of the optimal formamide concentration. Where available, pure cultures were obtained from DSMZ and applied in the optimization process. *Tetrasphaera japonica* (DSM13192) was used to optimize the probe Tetra183, while *Sanguibacter suarezii* (DSM10543), *Lactobacillus reuteri* (DSM20016), and *Janibacter melonis* (DSM16063) were used to assess the need for the specific unlabelled competitor probes Tetra67_C1, Actino221_C3, and Tetra732_C1, respectively. If appropriate pure cultures were not available, probes were optimized using activated sludge biomass with a high abundance of the target organism predicted by amplicon sequencing.

Raman microspectroscopy

Raman microspectroscopy was applied in combination with FISH, as previously described[38]. The approach was used to identify phenotypic differences between probe-defined *Tetrasphaera* phylotypes. Briefly, FISH was conducted on optically polished CaF_2 Raman windows (Crystran, UK), which give a single-sharp Raman marker at 321 cm^{-1} that serves as an internal reference point in every spectrum. *Tetrasphaera* species-specific (Cy3) probes (**Table S4**) were used to locate the target cells for Raman analysis. After bleaching the Cy3 fluorophore with the Raman laser, spectra from single cells were obtained using a Horiba LabRam HR 800 Evolution (Jobin Yvon – France) equipped with a Torus MPC 3000 (UK) 532 nm 341 mW solid-state semiconductor laser. The Raman spectrometer was calibrated prior to obtaining all measurements to the first-order Raman signal of Silicon, occurring at 520.7 cm^{-1}. The incident laser power density on the sample was attenuated down to 2.1 $mW/\mu m^2$ using a set of neutral density (ND) filters. The Raman system is equipped with an in-built Olympus (model BX-41) fluorescence microscope. A 50X, 0.75 numerical aperture dry objective (Olympus M Plan Achromat- Japan), with a working distance of 0.38 mm, was used throughout the work. A diffraction grating of 600 mm/groove was used, and the Raman spectra collected spanned the wavenumber region of 200 cm^{-1} to 1800 cm^{-1}. The slit width of the Raman spectrometer and the confocal pinhole diameter were set to 100 µm and 72 µm, respectively. Raman spectrometer operation and subsequent processing of spectra were conducted using LabSpec version 6.4 software (Horiba Scientific, France). All spectra were baseline corrected using a 6th order polynomial fit.

Results and discussion:

A comprehensive ecosystem-specific 16S rRNA gene reference database

In order to make a comprehensive ecosystem-specific reference database for Danish wastewater treatment plants (WWTPs) and their anaerobic digesters, we sampled biomass from 22 WWTPs and 16 anaerobic digesters (ADs) treating waste activated sludge located at Danish wastewater treatment facilities, all representative for Danish treatment facilities (**Table S2**). These facilities represent an important engineered ecosystem containing both bacterial and archaeal complex communities, with the vast majority of microbes being uncultured and poorly characterized[39].

DNA and RNA were extracted and pooled separately for each environment and used to create ecosystem-specific primer-based (DNA-based) and "primer-free" (RNA-based) fSSU libraries (**Figure 1a**). This resulted in a total of 926,507 fSSU sequences after quality filtering. A comprehensive reference database was constructed by accepting only sequences observed at least twice. We refer to these sequences as exact sequence variants (ESVs). As each fSSU is independently amplified due to the unique molecular identifiers (UMIs) added before the PCR amplification steps, the risk of having multiple ESVs with identical errors is extremely low if we assume random distribution of errors (see supplementary results). ESVs are therefore considered to be essentially error-free. The final ESV database contained 21,039 unique full-length rRNA gene sequences.

To determine the influence of library preparation method, we compared ESVs created based on fSSU obtained from the four individual libraries. The DNA-based approach yielded approx. 20 times more unique ESVs than the RNA-based approach for the same sequencing cost (**Table S3**). The reduced number of unique ESVs from the RNA-based libraries was expected, as only 13.3% of the assembled sequences represented full-length 16S rRNA gene sequences (**Table S3**). As the Archaea are not targeted by the primers used, we compared the bacterial ESVs from the four libraries to assess the influence of primer bias (**Figure 1b)**. This revealed that 27% and 32% of the unique ESVs identified in the shallow RNA-based libraries were not present in corresponding DNA-based libraries for activated sludge and anaerobic digesters, respectively. This reveals a clear bias associated with the DNA-based method, which is in accordance with our previous *in silico* evaluation of primer bias for the 27F and 1492R primer pair[17]. The same analysis predicted that a better coverage could be achieved by using the 27F and 1391R primer pair[40] on the expense of sequence length[17].

To estimate the number of ESVs belonging to novel taxa, these were mapped to the SILVA_132_SSURef_Nr99 database[12], and the identity of the closest relative was compared to the thresholds for taxonomic ranks proposed by Yarza *et al.*[9] (**Table 1**). The majority of the ESVs (~96%) had references in the SILVA database with genus-level support (identity >94.5%), but 20% lacked references above the species-level (identity > 98.7%) (**Table 1**), which are crucial to confident taxonomic classification[10].

Evaluation of the ESV database using amplicon data

In order to evaluate if the ESV database contained high-identity references for all prokaryotes in the ecosystem, we mapped V1-3 amplicon sequencing data obtained from two sources: the same samples used to create the ESV database and samples from unrelated Danish WWTP and ADs. To ensure maximal resolution, amplicon data was processed into ASVs. The ecosystem-specific ESV database (21,039 seq.) included more high-identity references for all analyzed samples, compared to 25-150-fold larger universal databases, such as MiDAS 2.1 (548,447 seq.)[21], SILVA v.132 SSURef Nr99 (695,171 seq.), SILVA v.132 SSURef (2,090,668 seq.), GreenGenes 16S v.13.5 (1,262,986 seq.), and the full RDP v.11.5 (3,356,808 seq.) (**Figure 1c and Figure S1-S2**). A decrease in percentage of ASVs with high-identity references was observed when ASVs with lower abundance were included in the analysis. However, the ESV database still performed as well as the larger universal databases.

Since only Danish WWTPs and ADs were used to establish the comprehensive high-identity ESV reference database, published amplicon data from non-Danish WWTPs[41,42] was also evaluated (**Figure 1d-e, and S3-S4**). Compared to all universal reference databases, the Danish reference ESVs performed better or similar for most of the investigated non-Danish WWTPs, although not as well as for the Danish plants, further demonstrating the advantage of using ecosystems-specific databases. The inclusion of sequences from non-Danish WWTP and ADs will likely improve classification for plants globally.

A new comprehensive taxonomic framework

A major limitation in the classification of amplicon data from environmental samples is the lack of lower rank taxonomic information (family, genus, and species names) for many uncultivated bacteria in the universal reference databases. To address this, we developed a robust taxonomic framework (AutoTax), which provides consistent taxonomic classification of all sequences to all seven taxonomic ranks by using a reproducible computational approach, based on identity thresholds (**Figure 2**).

The ESVs were first mapped to the SILVA_SSURef_Nr99 database, which provides the taxonomy of the closest relative in the database as well as the identity between the ESV and this reference. The taxonomy was assigned to the ESV down to the taxonomic rank that is supported by the sequence identity thresholds proposed by Yarza *et al.*[9]. As the SILVA taxonomy does not include species names, ESVs were also mapped to 16S rRNA gene sequences from type strains extracted from the SILVA database. Species names were added to the ESVs if the identity was above 98.7%, and the genus name obtained from the type strains was identical to that obtained from the full reference database. Although there are examples of separate species with 16S rRNA genes that share more than 98.7% sequence similarity and genomes with intragenomic copies that are less than 98.7% similar, these are exceptions rather than the norm[43,44]. The approach used here will therefore provide correct species-level classifications for the vast majority of the ESVs.

To fill gaps in the taxonomy, all ESVs were trimmed and clustered using the UCLUST cluster_smallmem algorithm and the taxonomic thresholds proposed by Yarza *et al.*[9]. With this algorithm sequences are processed in the order they appear in the input file, i.e., if the next sequence matches an existing centroid, it is assigned to that cluster, otherwise it becomes the centroid of a new cluster. This ensures that the same clusters and centroids are formed every time, even if additional ESVs are appended to the reference database. The reproducibility of the approach was confirmed by processing only the first half of the ESVs, which yielded identical clusters. Merging of the SILVA- and the *de novo*-based taxonomies may result in conflicts (e.g., multiple ESVs from the same species associate with different genera). When this is the case, the taxonomy for the ESV, which first appears in the reference database, is adapted for all ESVs within that species. The pipeline produces formatted reference databases, which can be directly used for classification using sintax or classifiers in the qiime2 framework.

AutoTax provided placeholder names for many previously undescribed taxa (**Table 2, Figure S5**). Strikingly, essentially all species, more than 70% of all genera, 50% of all families, and 30% of all orders obtained their names from the *de novo* taxonomy and would otherwise have remained unclassified. The novel taxa were affiliated with several phyla, especially the Proteobacteria, Planctomycetes, Patescibacteria, Firmicutes, Chloroflexi, Bacteroidetes, Actinobacteria, and Acidobacteria (**Figure S5**). A prominent example is the Chloroflexi, where only 10/14 orders, 8/33 families, and 10/151 genera observed here were classified using the SILVA database, clearly showing the need for an improved taxonomy. This will have important implications for the study of these communities, given the high diversity and abundance of members of this phylum and their association with the sometimes serious operational problems of bulking and foaming[45,46].

To benchmark the ESV database, we classified amplicon data obtained from activated sludge and anaerobic digester samples using this database and compared the results to classifications obtained from the universal reference databases (**Figure 3a**). The ESV database was able to classify many more of the ASVs to the genus level (~90%), compared to MiDAS 2.1 (~65%), SILVA_132_SSURef_Nr99 (~45%), GreenGenes_16s_13.5 (25-30%), and the RDP_16S_v16 training set (25%). Importantly, many of the top 50 most abundant ASVs only received classification with the ESV database (**Figure 4 and S6**). The use of the ESV database thus significantly improved the classification at all taxonomic levels and, importantly, provided species-level classifications for the majority of the ASVs (~85%).

Confident classification of amplicon sequences based on reference databases can be difficult due to the limited taxonomic information in short sequences[9,10]. To investigate the confidence of the amplicon classification, we extracted amplicon sequence sets *in silico* from the ESVs, corresponding to commonly amplified 16S rRNA regions. These amplicons were classified using sintax and the full-length ESV database. We then calculated the fraction of amplicons, which was correctly classified to the same genus and species as their source ESV (**Figure 3b**). Nearly 100% of the amplicons were assigned to the correct genus and 86-96% to the correct species, depending on primer set used to trim the sequences. The primers targeting the V1-3 variable region performed especially well for species-level identification (96% correct classifications), while the commonly used primers targeting the V4 variable region were among the worst (89% correct classifications). Similar levels of classification were also obtained when only the trimmed forward reads (250 bp) were used, compared to the merged forward and reverse reads (95.93% and 95.87% correct species-level classification, respectively; **Figure 3b**). This demonstrated that the use of a comprehensive, high-quality database allows the confident classification of ASV sequences to the genus to species level.

When choosing primers for amplicon sequence analyses, it is important also to take primer-bias into account[24]. If a poor choice is made, process-relevant species may not appear, or they may be severely underestimated. For activated sludge, it has previously been shown that the V1-3 primers have a good overall agreement with metagenomic data and capture many of the process-relevant organisms, whereas the V4 primers underestimate the abundance of the process-critical Chloroflexi and Actinobacteria[24]. Access to a comprehensive ecosystem-specific full-length 16S rRNA gene database provides an opportunity to determine the theoretical coverage of different primer sets *in silico* for the given ecosystem so that an informed decision can be made[40,47].

Species-specific FISH probes for* Tetrasphaera *spp.

A valuable benefit for the generation of ecosystem-specific databases is the design and selection of probes and primers for specific populations. FISH-based visualization of populations is central to many studies in microbial ecology, yet with the expanding 16S rRNA gene databases, finding probe sites allowing confident differentiation of target lineages is becoming increasingly difficult. Probe specificity and coverage are routinely assessed, based on all the sequences in public databases, yet both parameters may be very different when considering only the microorganisms present in the ecosystem of study. The use of ecosystem-specific databases therefore provides a more accurate assessment of probe specificity and coverage and will likely also allow the confident design and application of probes for targeting lineages at a higher taxonomic resolution, such as species.

To illustrate the benefit of using the new high-quality reference ESV database, more detailed analyses of the genus *Tetrasphaera* were performed. It is the most abundant genus in Danish WWTPs[21] and is associated with the polyphosphate-accumulating organism (PAO) phenotype, important for the capture and removal of phosphorus in the WWTPs[38,48,49]. Despite the importance of the genus, it is unknown how many species co-exist in these systems and whether they all possess the PAO metabolism. Phylogenetic analysis of 722 ESVs belonging to the genus *Tetrasphaera* retrieved in this study revealed an evident separation into 18 species across 22 Danish WWTPs, providing for the first time a comprehensive overview of the diversity of *Tetrasphaera* in activated sludge systems (**Figure 5a)**. Several of the retrieved sequences are identical to those of the described pure cultures, while the majority are novel and not present in existing databases. The 10 most abundant species are shown in **Figure 5b**. In order to reveal possible variations in morphology and physiology of *Tetrasphaera*, the new ESV database was used to design a comprehensive set of FISH probes covering the abundant species (**Figure 5a**). Of those, only the two most abundant species in Danish WWTPs were targeted by the existing probes (Actino-658 and Actino-221)[50] with high specificity and coverage. Other existing FISH probes targeting genus *Tetrasphaera*[48] did not show *in silico* high specificity and/or coverage. The new species-specific probes designed to target the remaining abundant species, which can create up to 2-3% of the biomass in some plants (**Figure 5a**), revealed different morphologies (rod-shaped cells, tetrads, filaments, **Figure 5c)**. Having probes for these different species most importantly allows *in situ* single cell analyses for each. Using these FISH probes in combination with Raman microspectroscopy, it was confirmed that all the FISH-defined *Tetrasphaera* species were likely PAOs, based on the presence of a large peak for poly-P (1170 cm^{-1}, **Figure 5d**). No Raman peaks were found for other intracellular storage compounds such as glycogen, PHA, or trehalose – consistent with current models for the physiology of the genus in these systems. Additionally, the new reference database was used to design a probe set (Tetra183 + Tetra617) for genus-level screenings of all abundant species of *Tetrasphaera* in Danish plants for (**Figure 5a**), which was otherwise not possible.

Concluding remarks

The current study demonstrates how high-throughput full-length 16S rRNA gene sequencing can be combined with a sequence identity-based approach to automatically establish near-complete ecosystems-specific reference databases. These databases were shown to be superior to the much larger public versions for microbial community analyses, including their use for amplicon sequencing and FISH analyses.

The comprehensive taxonomy and high-identity reference sequences for all abundant microbes from the selected ecosystem greatly improve the specificity of taxonomic classification of amplicons to the sub-genus to sub-species level. It also provides the design and confident use of FISH probes that can be used to illuminate the morphological and functional diversity at sub-genus level *in situ*, as exemplified here for the genus *Tetrasphaera*. The assessment of probe specificity and coverage is more meaningful using the smaller ecosystem-specific database, compared to the much larger, but broader, public versions, given the possible inclusion of poor-quality sequences and those from microorganisms not observed in the system studied in the latter.

The sequence identity-based approach for automated taxonomy assignment (AutoTax) represents a simple, cost effective strategy to provide a comprehensive taxonomy for all seven major taxonomic ranks within a short time frame (a few hours for the database created here). The use of a large curated database (SILVA) as the backbone of the taxonomy assignment ensures that the primary classification is in accordance with the current consensus within microbial taxonomy. As, with time, ESVs are introduced into the public databases and their taxonomies are manually curated, this will further improve the ESV classifications by AutoTax. The assignment of placeholder names for unclassified taxa based on sequence clustering provides stable reference names, which can act as surrogates until the taxa receive true taxonomic classifications. Although the sequence similarity clustering does not necessarily reflect the true evolutionary history of the microbes or the phenotypic characteristics, it still provides clusters that are of similar diversity to true taxa at the same taxonomic rank.

The AutoTax pipeline was optimized here for use with the SILVA database, but can easily be adapted for use with other reference databases. An interesting example could be the rapidly expanding genome taxonomy database (GTDB), which provides a standardized bacterial taxonomy for all taxonomic ranks, including the species level, based on genome phylogeny of single copy marker genes extracted from metagenome assembled genomes (MAGs)[51]. The approach can also be applied to provide placeholder taxonomies for gaps in the large public database, such as SILVA, GreenGenes, and RDP; noting that this will improve the classification rate for sequences, but will probably not provide the same resolution as the ecosystem-specific databases.

We believe the presented approach will have profound impact on the future of microbial community analyses in many fields, including the water sector, soil microbiology, and human health. The robust taxonomic framework will provide a common language for scientific communities, which we anticipate will ease the sharing of microbial knowledge and provide a platform for linking microbial identity with biological functions.

Acknowledgements

The project has been funded by the Danish Research Council (grant 6111-00617A), the Villum foundation (Dark Matter and grant 13351), and 25 wastewater treatment plants.

Conflict of interest:

The authors declare no conflict of interest.

References:

1. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science (80-.).* **350**, aac9823-18 (2015).
2. Single-, D. R. R. & Sequence, N. C. Deblur Rapidly Resolves Single-. **2**, 1–7 (2017).
3. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
4. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 81257 (2016). doi:10.1101/081257
5. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.* (2010). doi:10.1073/pnas.1000080107
6. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
7. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
8. Mcdonald, D. *et al.* American Gut : an Open Platform for Citizen Science. *mSystems* **3**, 1–28 (2018).
9. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
10. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**, e4652 (2018).
11. Desantis, T. Z. *et al.* Greengenes , a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. **72**, 5069–5072 (2006).
12. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-6 (2013).
13. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633-42 (2014).
14. Ritari, J., Salojärvi, J., Lahti, L. & de Vos, W. M. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* **16**, 1056 (2015).
15. Segota, I. & Long, T. A high-resolution pipeline for 16S-sequencing identifies bacterial strains in human microbiome. (2019).
16. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford).* **2010**, baq013 (2010).
17. Karst, S. M. *et al.* Retrieval of a million high-quality , full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **36**, 190–195 (2018).
18. Burke, C. M. & Darling, A. E. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ* **4**, e2492 (2016).
19. Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *bioRxiv* 392332 (2019). doi:10.1101/392332

20. Karst, S. M., Ziels, R. M., Kirkegaard, R. H. & Albertsen, M. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing. *bioRxiv* 645903 (2019). doi:10.1101/645903

21. McIlroy, S. J. *et al.* MiDAS 2.0: An ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. *Database* **2017**, 1–9 (2017).

22. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. *A Guide to the Natural History of Freshwater Lake Bacteria*. *Microbiology and Molecular Biology Reviews* **75**, (2011).

23. Glöckner, F. O. *et al.* 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J. Biotechnol.* **261**, 169–176 (2017).

24. Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H. & Nielsen, P. H. Back to basics - the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS One* **10**, e0132783 (2015).

25. Kirkegaard, R. H. *et al.* The impact of immigration on microbial community composition in full-scale anaerobic digesters. *Sci. Rep.* **7**, (2017).

26. Lane, D. J. 16S/23S rRNA sequencing. in *Nucleic Acid Techniques in Bacterial Systematics* (eds. Stackebrandt, E. & Goodfellow, M.) 115–175 (John Wiley and Sons, 1991). doi:10.1007/s00227-012-2133-0

27. Muyzer G Uitterlinden AG., de W. E. C. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *AEM* **59**, 695–700 (1993).

28. R Core Team. R: A language and environment for statistical computing. (2016).

29. RStudio Team. RStudio: Integrated Development Environment for R. (2015).

30. Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis*. *Springer* (Springer Science & Business Media, 2009). doi:10.1007/978-0-387-98141-3

31. Andersen, K. S. S., Kirkegaard, R. H., Karst, S. M. & Albertsen, M. ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* 299537 (2018). doi:10.1101/299537

32. Daims, H., Stoecker, K. & Wagner, M. Fluorescence in situ hybridization for the detection of prokaryotes. *Mol. Microb. Ecol.* **213**, 239 (2005).

33. Amann, R. I. *et al.* Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Env. Microbiol* **56**, 1919–1925 (1990).

34. Daims, H., Brühl, A., Amann, R., Schleifer, K. H. & Wagner, M. The domain-specific probe EUB338 is insufficient for the detection of all Bacteria: development and evaluation of a more comprehensive probe set. *Syst Appl Microbiol* **22**, 434–444 (1999).

35. Wallner, G., Amann, R. & Beisker, W. Optimizing fluorescent *in situ* hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. *Cytometry* **14**, 136–143 (1993).

36. Ludwig, W. *et al.* ARB: A software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).

37. Yilmaz, L. S., Parnerkar, S. & Noguera, D. R. MathFISH, a web tool that uses thermodynamics-based mathematical models for *in silico* evaluation of

oligonucleotide probes for fluorescence *in situ* hybridization. *Appl. Environ. Microbiol.* **77**, 1118–1122 (2011).

38. Fernando, E. Y. *et al.* Resolving the individual contribution of key microbial populations to enhanced biological phosphorus removal with Raman-FISH. *ISME J.* (2019). doi:10.1101/387795

39. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol.* (2019). doi:10.1038/s41564-019-0426-5

40. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, 1–11 (2013).

41. Isazadeh, S., Jauffur, S. & Frigon, D. Bacterial community assembly in activated sludge: mapping beta diversity across environmental variables. *Microbiologyopen* **5**, 1050–1060 (2016).

42. Gonzalez-Martinez, A. *et al.* Comparison of bacterial communities of conventional and A-stage activated sludge systems. *Sci. Rep.* **6**, (2016).

43. Kim, M., Oh, H.-S. S., Park, S.-C. C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**, 346–351 (2014).

44. Větrovský, T. & Baldrian, P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS One* **8**, 1–10 (2013).

45. McIlroy, S. J. S. J. *et al.* Genomic and in situ investigations of the novel uncultured Chloroflexi associated with 0092 morphotype filamentous bulking in activated sludge. *ISME J* **(In press)**, 1–12 (2016).

46. Petriglieri, F., Nierychlo, M., Nielsen, P. H., Jon, S. & Id, M. In situ visualisation of the abundant Chloroflexi populations in full-scale anaerobic digesters and the fate of immigrating species. 1–14 (2018).

47. Walters, W. A. *et al.* PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**, 1159–1161 (2011).

48. Nguyen, H. T. T., Le, V. Q., Hansen, A. A., Nielsen, J. L. & Nielsen, P. H. High diversity and abundance of putative polyphosphate-accumulating Tetrasphaera-related bacteria in activated sludge systems. *FEMS Microbiol. Ecol.* **76**, 256–267 (2011).

49. Marques, R. *et al.* Metabolism and ecological niche of Tetrasphaera and Ca. Accumulibacter in enhanced biological phosphorus removal. *Water Res.* **122**, 159–171 (2017).

50. Kong, Y., Nielsen, J. L. & Nielsen, P. H. Identity and Ecophysiology of Uncultured Actinobacterial Polyphosphate-Accumulating Organisms in Full-Scale Enhanced Biological Phosphorus Removal Plants. *Appl. Environ. Microbiol.* **71**, 4076 LP – 4085 (2005).

51. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).

52. Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Res.* **19**, 2317–

2323 (2009).

53. Apprill, A., Mcnally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **75**, 129–137 (2015).

54. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**, 1403–1414 (2016).

Figures:

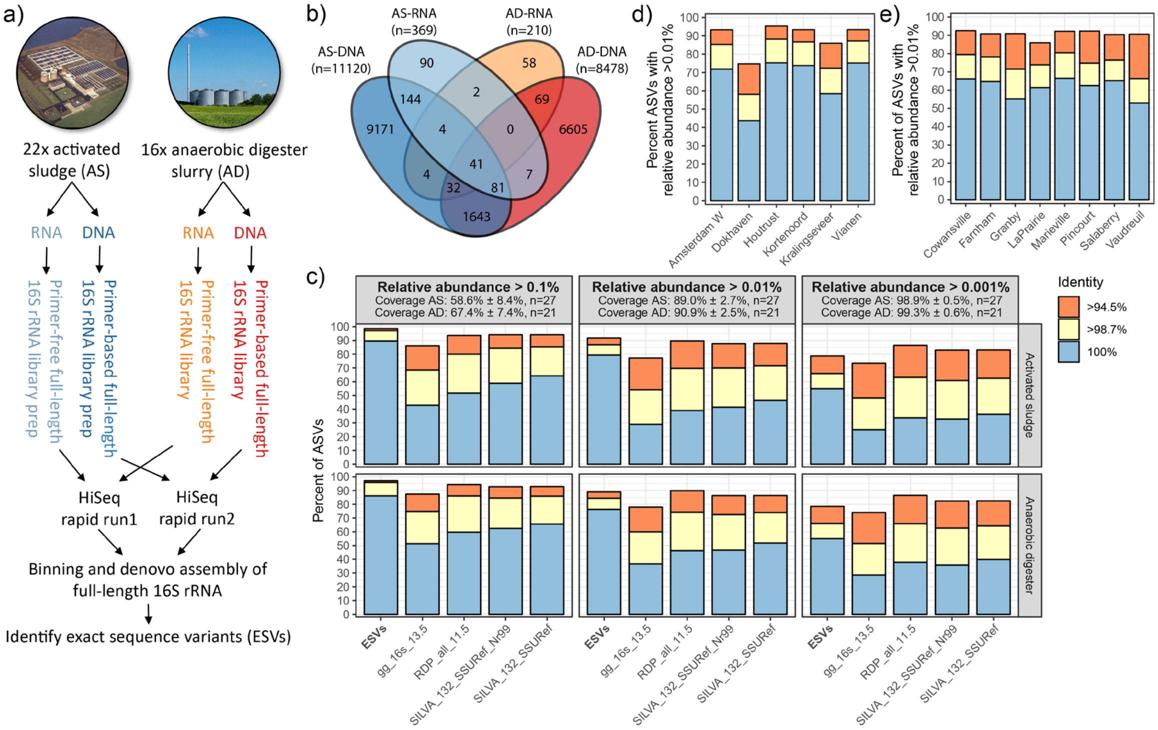

Figure 1. Construction and evaluation of the ESV reference database. a) Preparation of ESVs. Samples were collected from WWTPs and anaerobic digesters, and DNA and RNA were extracted. Purified DNA or RNA were used for preparation of primer-based and "primer-free" full-length 16S rRNA libraries, respectively. These were sequenced and processed bioinformatically to produce a comprehensive ecosystem-specific full-length 16S rRNA ESV database. A detailed description is provided in the supplementary results. b) Venn-diagram showing bacterial ESVs shared between individual libraries. c) Mapping of V1-3 amplicon data to the ESV database and common 16S rRNA reference databases. ASVs were obtained from activated sludge and anaerobic digester samples, and ASVs were filtered, based on their relative abundance, before the analysis to uncover the depth of the ESV database. The fraction of the microbial community remaining after the filtering (coverage) is shown as mean ± standard deviation across plants. d) Mapping of ASVs from Dutch WWTPs based on raw data from Gonzarlez-Martinez *et al.* [42]. For details, see Figure S3. e) Mapping of ASVs from Canadian WWTPs, based on raw data from Isazadeh *et al.* 2016[41]. For details, see Figure S4.

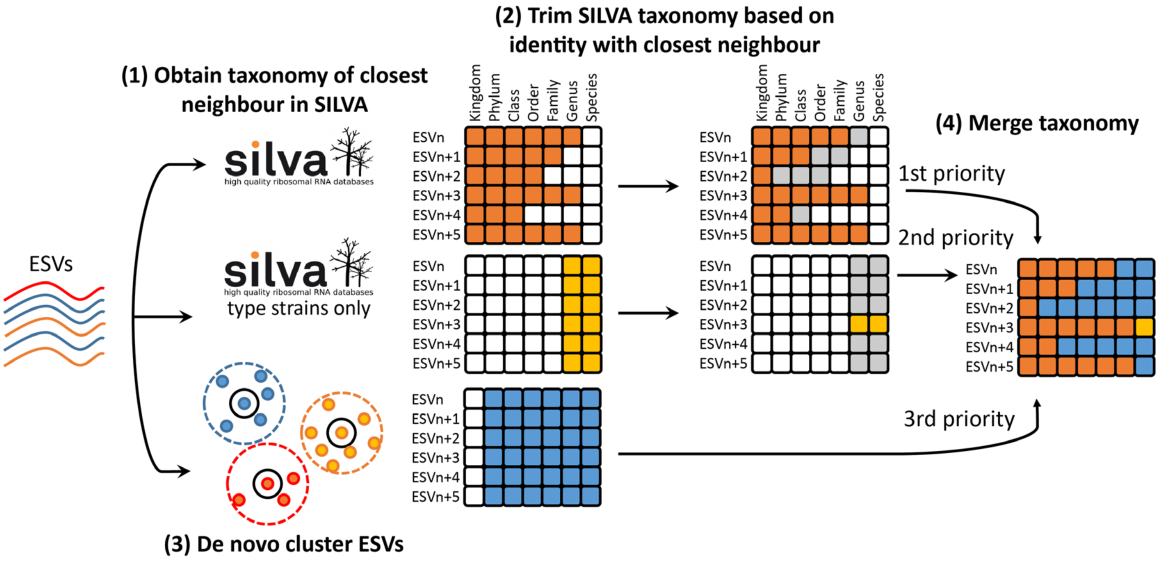
(1) Obtain taxonomy of closest neighbour in SILVA
(2) Trim SILVA taxonomy based on identity with closest neighbour
(3) De novo cluster ESVs
(4) Merge taxonomy
ESVs
silva high quality ribosomal RNA databases
silva high quality ribosomal RNA databases type strains only
Kingdom
Phylum
Class
Order
Family
Genus
Species
ESVn
ESVn+1
ESVn+2
ESVn+3
ESVn+4
ESVn+5
1st priority
2nd priority
3rd priority

Figure 2. The AutoTax taxonomic framework. (1) ESVs were first mapped to the SILVA_132_SSURef_Nr99 database to identify the closest relative. (2) Taxonomy was adopted from this sequence after trimming, based on identity and the taxonomic thresholds proposed by Yarza *et al.*[9]. To gain species information, ESVs were also mapped to sequences from type strains extracted from the SILVA database, and species names were adopted if the identity was >98.7% and the type strain genus matched that of the closest relative in the complete database. (3) ESVs were also clustered by greedy clustering at different identities, corresponding to the thresholds proposed by Yarza *et al.*[9] to generate a stable *de novo* taxonomy. (4) Finally, a comprehensive taxonomy was obtained by filling gaps in the SILVA-based taxonomy with the *de novo*-taxonomy.

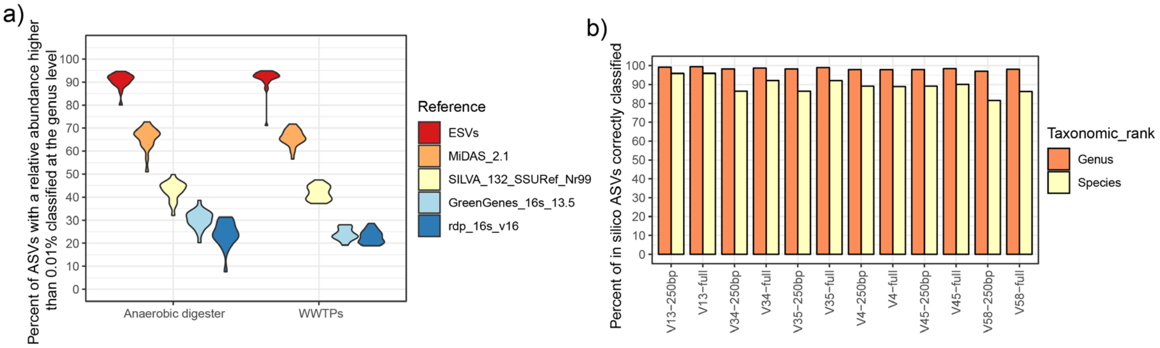

Figure 3. Classification of amplicons. a) Percentage of ASVs from each activated sludge and anaerobic digester sample with a relative coverage of more than 0.01% that were classified to the genus level when classified using the ESV reference database or common reference databases for taxonomic classification. b) Classification of *in silico* ASVs, corresponding to amplicons produced using common primer set on the ESVs. Results are shown for the full amplicons as well as for partial amplicons, equivalent to 250 bp forward reads of the amplicons. V13 (Lane 1991)[26], V34 (Klindworth et al. 2013)[40], V35 (Peterson et al. 2009)[52], V4 (Apprill et al. 2014)[53], V45 (Parada et al. 2015)[54], and V58 (Klindworth et al. 2013)[40].

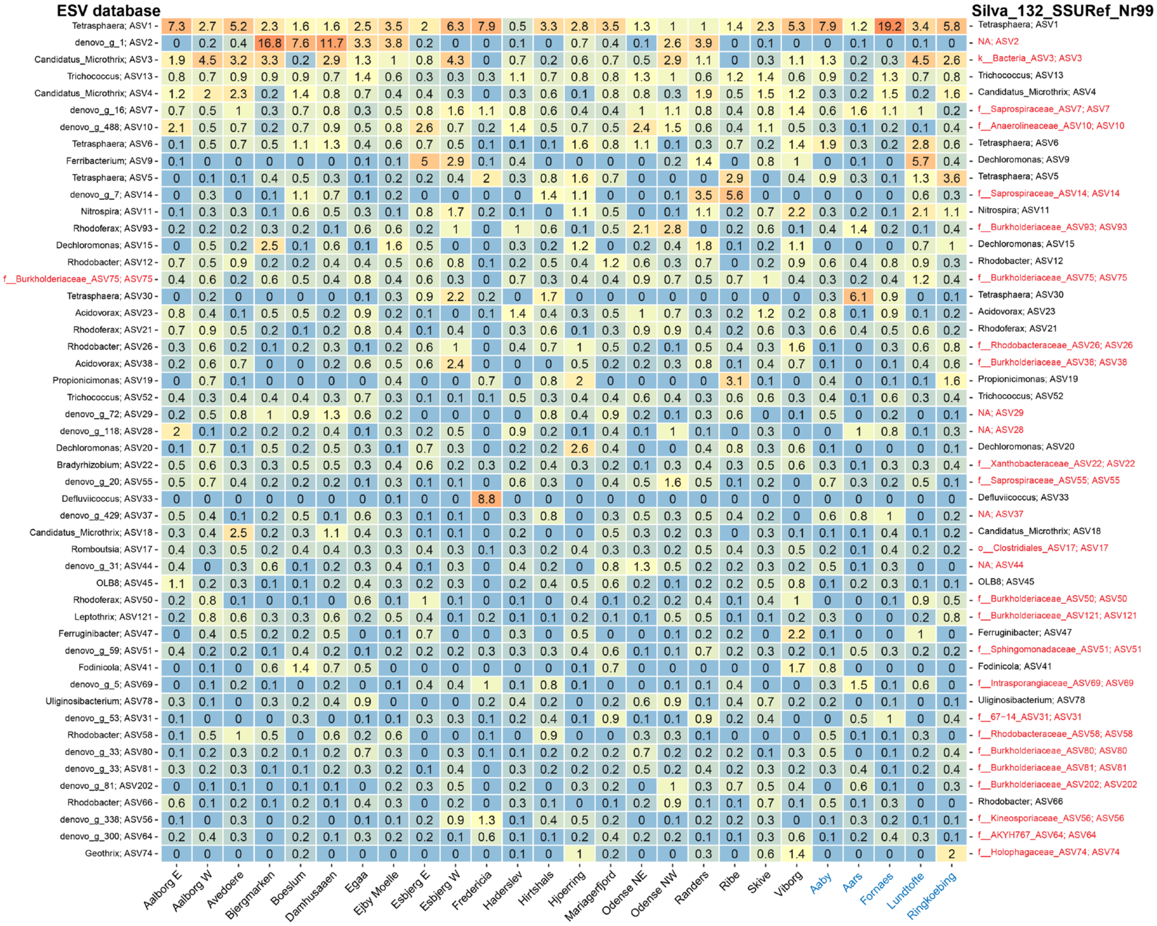

Figure 4. Relative abundance of the top 50 ASVs in the activated sludge samples

ASVs obtained from individual activated sludge samples used to create ESV database (black labels) and from other plants (blue labels) were classified based on the ESV database as well as the SILVA_132_SSURef_Nr99 database. ASVs not classified at the genus level have been highlighted in red.

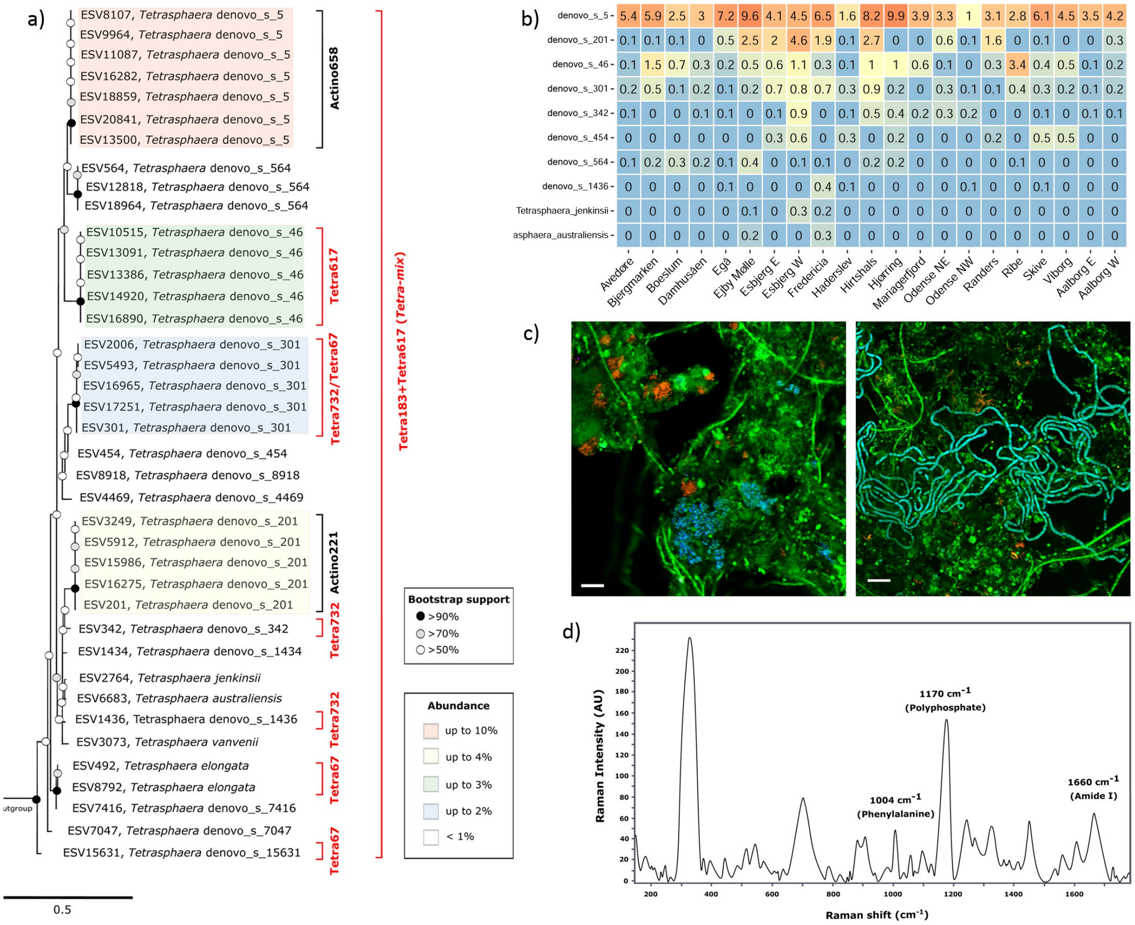

Figure 5. Detailed investigation into the genus *Tetrasphaera*. a) Maximum-likelihood (PhyML) 16S rRNA gene phylogenetic tree of activated sludge *Tetrasphaera* species and coverage of the existing (black) and new FISH probes (red). A 20% conservational filter was applied to the alignment used for the tree to remove hypervariable regions, giving 1422 positions. Coverage of probes relevant to the current study is shown in black brackets. Bootstrap values from 1000 re-samplings are indicated for branches with >50% (white dot), 50%–70% (gray), and >90% (black) support. Species of the genus *Dechloromonas* were used as the outgroup. The scale bar represents substitutions per nucleotide base. b) Abundance of top 10 *Tetrasphaera* species in full-scale activated sludge WWTPs sampled 2–4 times per year from 2006 to 2018. c) Composite FISH micrographs of chosen *Tetrasphaera* species from full-scale activated sludge WWTP. Left panel: rods of denovo_s_46 (orange) and tetrads of denovo_s_201 (blue) targeted by the probe Tetra617 and Actino221[50], respectively. Right panel: filaments of denovo_s_342 (cyan) and rods of denovo_s_5 (orange) targeted by the probe Actino658[50] and Tetra732, respectively. In both images, all other bacteria (green) are targeted by the probe EUBmix. Scale bars represent 10 µm. d) Raman spectrum of *Tetrasphaera* denovo_s_46 cells targeted by the Tetra617 probe. The presence of the signature peak for polyphosphate (1170 cm^{-1}) indicates the

potential accumulation of polyphosphate as intracellular storage compound. Peaks for phenylalanine (1004 cm^{-1}) and amide I linkages of proteins (1450 cm^{-1}), highlighted in the spectrum, are specific markers for biological material.

Tables:

Table 1: Numbers and percentage of ESVs estimated to belong to novel taxa. ESVs were mapped to SILVA_132_SSURef_Nr99 using usearch_global -id 0.5 –maxrejects 0 –strand plus to find the identity with the closest relative in the database. Novelty was determined, based on the identity for each ESV, based on the taxonomic thresholds proposed by Yarza et al. (2014) [9].

Environment	Library	Kingdom	nSeqs	Species [<98.7%]	Genus [<94.5%]	Family [<86.5%]	Order [<82.0%]	Class [<78.5%]	Phylum [<75.0%]
Activated sludge	RNA-based	Archaea	6	0	0	0	0	0	0
Activated sludge	RNA-based	Bacteria	499	43 / 8.61%	2 / 0.4%	0	0	0	0
Activated sludge	DNA-based	Bacteria	11266	2053 / 18.2%	398 / 3.53%	38 / 0.34%	8 / 0.071%	4 / 0.036%	3 / 0.027%
Anaerobic dig.	RNA-based	Archaea	274	10 / 3.65%	0	0	0	0	0
Anaerobic dig.	RNA-based	Bacteria	262	40 / 15.3%	9 / 3.44%	0	0	0	0
Anaerobic dig.	DNA-based	Bacteria	8638	2100 / 24.3%	399 / 4.62%	31 / 0.36%	9 / 0.10%	5 / 0.058%	3 / 0.035%

Table 2: Numbers and percentage of taxa which were assigned *de novo* names.

Environment	Library type	Kingdom	Species	Genus	Family	Order	Class	Phylum
Activated sludge	RNA-based	Archaea	2 / 100%	0	0	0	0	0
Activated sludge	RNA-based	Bacteria	196 / 92.5%	47 / 40.1%	14 / 20.0%	4 / 8.00%	0	0
Activated sludge	DNA-based	Bacteria	2745 / 94.8%	899 / 71.0%	181 / 44.1%	49 / 24.9%	8 / 10.0%	1 / 2.86%
Anaerobic digester	RNA-based	Archaea	12 / 70.6%	0	0	0	0	0
Anaerobic digester	RNA-based	Bacteria	117 / 91.4%	50 / 51.0%	21 / 30.4%	6 / 12.0%	3 / 9.68%	0
Anaerobic digester	DNA-based	Bacteria	1793 / 94.5%	605 / 66.4%	131 /43.0%	39 /23.1%	11 / 14.3%	0