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, 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

****Present address:** Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, 4072 Brisbane, Australia.

Running title:

Species-level resolution in microbial ecology

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 (ASVs) 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 genus to species rank (Martiny *et al.*, 2015).

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 known as sub-OTUs (sOTUs) or zero-radius OTUs (zOTUs), with denoising algorithms such as Deblur (Single- and Sequence, 2017), DADA2 (Callahan *et al.*, 2016), or Unoise3 (Edgar, 2016b). The sequences are finally classified, based on a 16S rRNA gene reference database to assign the most plausible taxonomy for each sequence (Caporaso *et al.*, 2010). 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 (Callahan *et al.*, 2017).

ASVs can be applied as consistent labels for microbial identification independently of a 16S rRNA gene reference database (Callahan *et al.*, 2017). This approach is used in several large-scale projects, including the Earth Microbiome Project (EMP) (Thompson *et al.*, 2017) and the American Gut project (Mcdonald *et al.*, 2018), 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 (Yarza *et al.*, 2014; Edgar, 2018), 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 and improved sequencing technologies that can produce longer reads of high quality. Finally, if information about functional properties is available from pure culture studies or *in situ* studies based on metagenome assembled genomes (MAGs), this information may be linked to full-length 16S rRNA sequences, but less reliably to ASVs (Yarza *et al.*, 2014; Edgar, 2018). 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 (Edgar, 2016a) or RDP classifier (Wang *et al.*, 2007)) 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 (Edgar, 2018). None of these criteria are met with the commonly applied universal reference databases (Greengenes (Desantis *et al.*, 2006), SILVA (Quast *et al.*, 2013), and RDP (Cole *et al.*, 2014)), 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 (Ritari *et al.*, 2015) or oral microbiomes (Chen *et al.*, 2010; Griffen *et al.*, 2011), now have fairly comprehensive reference databases with genus- to species-level resolution, which were obtained from thousands of isolates and MAGs (Ritari *et al.*, 2015; Segota and Long, 2019; Chen *et al.*, 2010). 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 (Karst *et al.*, 2018; Burke and Darling, 2016), but also emerging methods such as PacBio (Callahan *et al.*, 2019) and Nanopore (Karst *et al.*, 2019) 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. However, it 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 (McIlroy *et al.*, 2017) and the Dictyopteran gut microbiota reference Database (DictDb) (Mikaelyan *et al.*, 2015). Another approach is to develop 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 (Newton *et al.*, 2011; Rohwer *et al.*, 2018), the honey bee gut microbiota database (Newton and Roeselers, 2012), and the rumen and intestinal methanogen database (Seedorf *et al.*, 2014). Such ecosystem-specific databases greatly improve classification of amplicons that have a suitable reference in the database, but this may not be the case for a large fraction of the community. Furthermore, manual ecosystem-specific curation of the reference databases is subjective and hardly sustainable if we want to expand the databases so that they cover the true diversity of the ecosystems at high taxonomic resolution, which would probably require 100-1000 times more sequences (Glöckner *et al.*, 2017).

Ideally, we want an automated taxonomy assignment that can provide robust, objective taxonomic classifications for all 16S rRNA gene reference sequences, 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 (Yarza *et al.*, 2014). 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 version of the SILVA reference database.

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 9,521 full-length exact sequence variants (ESVs), which were classified using AutoTax. The value of the new approach was demonstrated by comparing the performance of the ESV database with the large universal reference database commonly applied. The comprehensive set of full-length 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 (McIlroy *et al.*, 2017). 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 (Albertsen *et al.*, 2015). 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 (Karst *et al.*, 2018). The SSU_rRNA_RT2 (activated sludge) and SSU_rRNA_RT3 (anaerobic digester biomass) reverse transcription primer and the SSU_rRNA_1 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 (Karst *et al.*, 2018).

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 (Karst *et al.*, 2018). 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 (Karst *et al.*, 2018).

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) (Karst *et al.*, 2018). 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 (Karst *et al.*, 2018). This ensures that the sequences have equal length and that primer binding sites are removed from the DNA-based sequences. The trimmed sequences were oriented according to the SILVA_132_SSURef_Nr99 database (Quast *et al.*, 2013) using the usearch11 -orient command, dereplicated using usearch11 -fastx_uniques -sizeout and denoised with usearch11 -unoise3 -minsize 2 to produce "error-free" full-length ESVs. For details see the supplementary results.

Taxonomy assignment to full-length ESVs

A complete taxonomy from kingdom to species was automatically assigned to each full-length 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 using usearch, 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*. (Yarza *et al.*, 2014). In addition, full-length ESVs were *de novo* clustered using the UCLUST algorithm and the same thresholds. The *de novo* clusters were labelled based on number of the centroid ESV, and these labels acted as a placeholder taxonomy, where there were gaps in the taxonomy obtained from SILVA. For details, see the supplementary results.

193 **Amplicon sequencing and analysis**

194 Bacterial community analysis was performed by amplicon sequencing of the V1-3 variable region
195 as previous described (Kirkegaard *et al.*, 2017) using the 27F (AGAGTTTGATCCTGGCTCAG)
196 (Lane, 1991) and 534R (ATTACCGCGGCTGCTGG) (Muyzer *et al.*, 1993) primers and the
197 purified DNA from above. Forward reads were processed using usearch v.11.0.667. Raw fastq files
198 were filtered for phiX sequences using -filter_phix, trimmed to 250 bp using -fastx_truncate -
199 trunclen 250, and quality filtered using -fastq_filter with -fastq_maxee 1.0. The sequences were
200 dereplicated using -fastx_uniques with -sizeout -relabel Uniq. Exact amplicon sequence variants
201 (ASVs) were generated using -unoise3 (Edgar, 2016b). ASV-tables were created by mapping the
202 raw reads to the ASVs using -otutab with the -zotus and -strand both options. Taxonomy was
203 assigned to ASVs using -sintax with -strand both and -sintax_cutoff 0.8 (Edgar, 2018).
204
205 **Data analysis and visualization**
206 Usearch v.11.0.667 was used for mapping sequences to references with -usearch_global -id 0 -
207 maxrejects 0 -maxaccepts 0 -top_hit_only -strand plus, unless otherwise stated. Data was imported
208 into R (R Core Team, 2016) using RStudio IDE (RStudio Team, 2015), analysed, and aggregated
209 using Tidyverse v.1.2.1 (https://www.tidyverse.org/), and visualised using ggplot2 (Wickham,
210 2009) v.3.1.0 and Ampvis (Andersen *et al.*, 2018) 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. The AutoTax script and processed reference databases in sintax and qiime format can be found at https://github.com/KasperSkytte/AutoTax.

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. (2005). Details about the optimal formamide concentration used for each probe are given in **Table S4**. The EUBmix probe set (Amann *et al.*, 1990; Daims *et al.*, 1999) was used to cover all bacteria, and the nonsense NON-EUB probe (Wallner *et al.*, 1993) 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 (Ludwig *et al.*, 2004). A phylogenetic tree was calculated, based on the aligned 72 new full-length 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 (Yilmaz *et al.*, 2011). 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 (Fernando *et al.*, 2019). 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 6^{th} 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 typical WWTPs and 16 anaerobic digesters (ADs) treating waste activated sludge located at Danish wastewater treatment facilities (**Table S2**). These facilities represent an important engineered ecosystem containing complex microbial communities of both bacteria and archaea, with the vast majority of microbes being uncultured and poorly characterized (Wu *et al.*, 2019).

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. The raw sequences were dereplicated and denoised with Unoise3 to generate a comprehensive reference database of 9,521 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). The ESVs are therefore considered to be essentially error-free.

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. 12 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 28% and 31% of the unique ESVs identified in the shallow RNA-based libraries were not present in corresponding

294 DNA-based libraries for activated sludge and anaerobic digesters, respectively. This reveals a bias
295 associated with the DNA-based method, which is in accordance with previous *in silico* evaluation
296 of primer bias for the 27F and 1492R primer pair (Karst *et al.*, 2018). The same study predicted that
297 a better coverage could be achieved by using the 27F and 1391R primer pair (Klindworth *et al.*,
298 2013) on the expense of sequence length (Karst *et al.*, 2018).

299 To estimate the number of full-length ESVs belonging to novel taxa, ESVs were mapped to
300 the SILVA_132_SSURef_Nr99 database (Quast *et al.*, 2013) using usearch, and the identity of the
301 closest relative was compared to the thresholds for taxonomic ranks proposed by Yarza *et al.* 2014)
302 (**Table 1**). The majority of the ESVs (~94%) had references in the SILVA database with genus-
303 level support (identity >94.5%), but 26% lacked references above the species-level (identity >
304 98.7%) (**Table 1**), which are crucial to confident taxonomic classification (Edgar, 2018).
305
306 ***Evaluation of the full-length ESV database using amplicon data***
307 In order to evaluate if the full-length ESV database contained high-identity references for all
308 prokaryotes in the ecosystem, we mapped V1-3 amplicon sequencing data obtained from two
309 sources: the same samples used to create the ESV database and samples from unrelated Danish
310 WWTP and ADs. To ensure highest resolution, amplicon data was processed into ASVs. The
311 ecosystem-specific ESV database (9,521 seq.) included more high-identity references (>98.7%
312 identity) for all analyzed samples, compared to the 58-353-fold larger universal databases, such as
313 MiDAS 2.1 (548,447 seq.)(McIlroy *et al.*, 2017), SILVA v.132 SSURef Nr99 (695,171 seq.),
314 SILVA v.132 SSURef (2,090,668 seq.), GreenGenes 16S v.13.5 (1,262,986 seq.), and the full RDP
315 v.11.5 (3,356,808 seq.) (**Figure 1c and Figure S1-S2**). A decrease in percentage of ASVs with
316 high-identity references was observed when low abundant ASVs (the rare biosphere) were included
317 in the analysis. However, the full-length ESV database still performed as well as the larger

318 universal databases. ASVs were also mapped to the 16S rRNA database derived from the Genome
319 Taxonomy Database (GTDB) release 89 (17,460 seq) (Parks *et al.*, 2018). However, this database
320 lacked high-identity references for almost all ASVs, which probably relates to the fact that 16S
321 rRNA genes often fail to assemble in MAGs produced by short read sequencing data.
322 Since only Danish WWTPs and ADs were used to establish the comprehensive high-identity
323 full-length ESV reference database, published amplicon data from non-Danish WWTPs (Isazadeh *et*
324 *al.*, 2016; Gonzalez-Martinez *et al.*, 2016) was also evaluated (**Figure 1d-e, and S3-S4**). Compared
325 to all universal reference databases, the Danish reference ESVs performed better or as well for most
326 of the investigated non-Danish WWTPs, which indicates that the database covers many of the
327 microbes that are common in WWTP across the world. We anticipate that high-throughput 16S
328 rRNA gene sequencing of non-Danish WWTP and ADs will provide references for the region-
329 specific taxa in the future.
330
331 ***A new comprehensive taxonomic framework***
332 A major limitation in the classification of amplicon data from environmental samples is the lack of
333 lower rank taxonomic information (family, genus, and species names) for many uncultivated
334 bacteria in the universal reference databases. To address this, we developed a robust taxonomic
335 framework (AutoTax), which provides consistent taxonomic classification of all sequences to all
336 seven taxonomic ranks by using a reproducible computational approach, based on identity
337 thresholds (**Figure 2**).
338 The full-length ESVs were first mapped to the SILVA_SSURef_Nr99 database, which
339 provides the taxonomy of the closest relative in the database as well as the percent identity between
340 the ESV and this reference. The taxonomy was assigned to the ESV down to the taxonomic rank
341 that is supported by the sequence identity thresholds proposed by Yarza *et al.* (2014) (**Table 1**). As

342 the SILVA taxonomy does not include species names, ESVs were also mapped to 16S rRNA gene
343 sequences from type strains extracted from the SILVA database. Species names were added to the
344 ESVs if the identity was above 98.7%, and the genus name obtained from the type strains was
345 identical to that obtained from the complete SILVA database. Although there are examples of
346 separate species with 16S rRNA genes that share more than 98.7% sequence similarity and
347 genomes with intragenomic copies that are less than 98.7% conserved, these are exceptions rather
348 than the norm (Kim *et al.*, 2014; Větrovský and Baldrian, 2013). The AutoTax approach will
349 therefore provide confident species-level classifications for the vast majority of the ESVs.
350 To fill gaps in the taxonomy, all ESVs were trimmed and clustered using the UCLUST
351 cluster_smallmem algorithm and the taxonomic thresholds proposed by Yarza *et al.* (2014). With
352 this algorithm sequences are processed in the order they appear in the input file, i.e., if the next
353 sequence matches an existing centroid, it is assigned to that cluster, otherwise it becomes the
354 centroid of a new cluster. This ensures that the same clusters and centroids are formed every time,
355 even if additional ESVs are added to the reference database in the future. The reproducibility of the
356 approach was confirmed by processing only the first half of the ESVs, which yielded identical
357 clusters. Merging of the SILVA- and the *de novo*-based taxonomies may result in conflicts (e.g.,
358 multiple ESVs from the same species associate with different genera). When this is the case, the
359 taxonomy for the ESV, which first appears in the reference database, is adapted for all ESVs within
360 that species. The pipeline produces formatted reference databases, which can be directly used for
361 classification using sintax or classifiers in the qiime2 framework.
362 AutoTax provided placeholder names for many previously undescribed taxa (**Table 2,**
363 **Figure S5**). Essentially all species, more than 72% of all genera, 50% of all families, and 30% of all
364 orders, obtained their names from the *de novo* taxonomy and would otherwise have remained
365 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 9/14 orders, 8/34 families, and 10/152 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 (McIlroy *et al.*, 2016; Petriglieri *et al.*, 2018).

To benchmark the full-length 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 SILVA_132_SSURef_Nr99 (~45%), GreenGenes_16s_13.5 (~25-30%), GTDB_r89 (~20-25%), and the RDP_16S_v16 training set (~25%) but also compared to gold standard within wastewater treatment systems MiDAS 2.1 (~65%), which is a manually ecosystem-specific curated version of the SILVA_123_SSURef_Nr99 database. Importantly, many of the top 50 most abundant ASVs only received classification with the AutoTax processed ESV database (**Figure 4 and S6**).

The use of the ecosystem-specific full-length ESV database thus significantly improved the classification at all taxonomic levels and, importantly, provided species-level classifications for the majority of the ASVs (~85%). To investigate the effect of the taxonomy assignment by AutoTax alone, we processed the SILVA_132_SSURef_Nr99 database using AutoTax. This increased the percentage of ASVs classified at the genus-level from ~45% to ~75%, and at the species-level from 0% to ~45%, demonstrating that the large universal databases would also benefit from the use of AutoTax. However, as better classifications are obtained using smaller ecosystem-specific

389 databases, we anticipate that such databases should be used whenever a defined ecosystem is
390 studied.

391 Confident classification of amplicon sequences based on reference databases can be
392 challenging due to the limited taxonomic information in short sequences (Yarza *et al.*, 2014; Edgar,
393 2018). To investigate the confidence of the amplicon classification, we extracted ASV sets *in silico*
394 from the bacterial full-length ESVs, corresponding to commonly amplified 16S rRNA regions.
395 These ASVs were classified using sintax and the full-length ESV database. We then calculated the
396 fraction of amplicons, which was correctly classified to the same genus and species as their source
397 ESV (**Figure 3b**). More than 95% of the ASVs were assigned to the correct genus and 72-90% to
398 the correct species, depending on primer set used. The primers targeting the V1-3 variable region
399 performed especially well for species-level identification (90% correct classifications), while the
400 commonly used primers targeting the V4 variable region were among the worst (72% correct
401 classifications). Further analysis of the ASV that did not receive a correct classification revealed
402 that the majority did not get any classification, and very few obtained wrong classifications (<0.2%
403 at genus-level and <0.8% at species-level).

404 Sequencing costs can be reduced considerably if single reads are used instead of merged
405 reads. To evaluate the effect of reduced amplicon length, trimmed forward reads (250 and 200 bp)
406 were also classified (**Figure 3b**). The percentage of correct classifications decreased only
407 marginally for the V1-3, V4, and V45 primer sets, whereas a more pronounced effect was observed
408 for the V3-4, V3-5, and V5-8 ASVs. Single reads of 250 or 200 bp are thus sufficient for some
409 primer sets.

410 Overall, the analysis demonstrated that the use of a comprehensive, high-quality reference
411 database allows the confident classification of ASV sequences at the genus to species level
412 depending of the primer sets used.

When choosing primers for amplicon sequence analyses, it is important also to take primer-bias into account (Albertsen *et al.*, 2015). 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 (Albertsen *et al.*, 2015). 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 (Walters *et al.*, 2011; Klindworth *et al.*, 2013).

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 (McIlroy *et al.*, 2017) and is associated with the polyphosphate-accumulating organism (PAO) phenotype, important for the capture and removal of phosphorus in the WWTPs (Nguyen *et*

437 *al.*, 2011; Marques *et al.*, 2017; Fernando *et al.*, 2019). Despite the importance of the genus, it is
438 unknown how many species co-exist in these systems and whether they all possess the PAO
439 metabolism. Phylogenetic analysis of 75 ESVs belonging to the genus *Tetrasphaera* retrieved in
440 this study revealed an evident separation into 19 species across 22 Danish WWTPs, providing for
441 the first time a comprehensive overview of the diversity of *Tetrasphaera* in activated sludge
442 systems (**Figure 5a)**. Several of the sequences retrieved were identical to those of the described
443 pure cultures, while the majority were novel and not present in existing databases. The 10 most
444 abundant species are shown in **Figure 5b**. In order to reveal possible variations in morphology and
445 physiology of *Tetrasphaera*, the new ESV database was used to design a comprehensive set of
446 FISH probes covering the abundant species (**Figure 5a**). Of those, only the two most abundant
447 species in Danish WWTPs were targeted by the existing probes (Actino-658 and Actino-221)
448 (Kong *et al.*, 2005) with high specificity and coverage. Other existing FISH probes targeting genus
449 *Tetrasphaera* (Nguyen *et al.*, 2011) did not show *in silico* high specificity and/or coverage.
450 The new species-specific probes designed to target the remaining abundant species, which
451 can create up to 2-3% of the biomass in some plants (**Figure 5a**), revealed different morphologies
452 (rod-shaped cells, tetrads, filaments, **Figure 5c)**. Having probes for these different species most
453 importantly allows *in situ* single cell analyses for each. Using these FISH probes in combination
454 with Raman microspectroscopy, it was confirmed that all the FISH-defined *Tetrasphaera* species
455 were likely PAOs, based on the presence of a large peak for poly-P (1170 cm^{-1}, **Figure 5d**). No
456 Raman peaks were found for other intracellular storage compounds such as glycogen, PHA, or
457 trehalose – consistent with current models for the physiology of the genus in these systems.
458 Additionally, the new reference database was used to design a probe set (Tetra183 + Tetra617) for
459 genus-level screenings of all abundant species of *Tetrasphaera* in Danish plants for (**Figure 5a**),
460 which was otherwise not possible..

Acknowledgements

We would like to thank the 25 wastewater treatment plants involved in the project for providing samples. This work was supported by the Danish Research Council [6111-00617A to P.H.N.]; and the Villum foundation [13351 to P.H.N.].

Conflict of Interest

The authors declare no conflict of interest.

468 **References:**

469 Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. (2015). Back to basics - the
470 influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge
471 communities. *PLoS One* **10**: e0132783.

472 Amann RI, Binder BJ, Olson RJ, Chisolm SW, Devereux R, Stahl DA. (1990). Combination of 16S
473 rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial
474 populations. *Appl Env Microbiol* **56**: 1919–1925.

475 Andersen KSS, Kirkegaard RH, Karst SM, Albertsen M. (2018). ampvis2: an R package to analyse
476 and visualise 16S rRNA amplicon data. *bioRxiv* 299537.

477 Apprill A, Mcnally S, Parsons R, Weber L. (2015). Minor revision to V4 region SSU rRNA 806R
478 gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol* **75**: 129–
479 137.

480 Burke CM, Darling AE. (2016). A method for high precision sequencing of near full-length 16S
481 rRNA genes on an Illumina MiSeq. *PeerJ* **4**: e2492.

482 Callahan BJ, McMurdie PJ, Holmes SP. (2017). Exact sequence variants should replace operational
483 taxonomic units in marker-gene data analysis. *ISME J* **11**: 2639–2643.

484 Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. (2016). DADA2:
485 High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

486 Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, *et al.* (2019). High-throughput
487 amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic*
488 *Acids Res* 1–12.

489 Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, *et al.* (2010).
490 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl*
491 *Acad Sci* **108**: 4516 LP – 4522.

492 Chen T, Yu W-H, Izard J, Baranova O V, Lakshmanan A, Dewhirst FE. (2010). The Human Oral
493 Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and
494 genomic information. *Database (Oxford)* **2010**: baq013.

495 Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, *et al.* (2014). Ribosomal Database
496 Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**: D633-42.

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

500 Daims H, Stoecker K, Wagner M. (2005). Fluorescence *in situ* hybridization for the detection of
501 prokaryotes. In: Osborn AM, Smith CJ (eds). *Molecular microbial ecology*. New York: Taylor &
502 Francis Group, pp 213–239.

503 Desantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006). Greengenes , a
504 Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. **72**: 5069–
505 5072.

506 Edgar R. (2016a). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences.
507 *bioRxiv* 074161.

508 Edgar RC. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences.
509 *PeerJ* **6**: e4652.

510 Edgar RC. (2016b). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon
511 sequencing. *bioRxiv* 81257.

512 Fernando EY, McIlroy SJ, Nierychlo M, Herbst F-A, Schmid MC, Wagner M, *et al.* (2019).
513 Resolving the individual contribution of key microbial populations to enhanced biological
514 phosphorus removal with Raman-FISH. *ISME J.* e-pub ahead of print, doi: 10.1101/387795.

515 Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, *et al.* (2017). 25 years of
516 serving the community with ribosomal RNA gene reference databases and tools. *J Biotechnol* **261**:
517 169–176.

518 Gonzalez-Martinez A, Rodriguez-Sanchez A, Lotti T, Garcia-Ruiz MJ, Osorio F, Gonzalez-Lopez
519 J, *et al.* (2016). Comparison of bacterial communities of conventional and A-stage activated sludge
520 systems. *Sci Rep* **6**. e-pub ahead of print, doi: 10.1038/srep18786.

521 Griffen AL, Beall CJ, Firestone ND, Gross EL, DiFranco JM, Hardman JH, *et al.* (2011). CORE: A
522 phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLoS One* **6**: 1–10.

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

525 Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. (2018). Retrieval
526 of a million high-quality , full-length microbial 16S and 18S rRNA gene sequences without primer
527 bias. *Nat Biotechnol* **36**: 190–195.

528 Karst SM, Ziels RM, Kirkegaard RH, Albertsen M. (2019). Enabling high-accuracy long-read
529 amplicon sequences using unique molecular identifiers and Nanopore sequencing. *bioRxiv* 645903.

530 Kim M, Oh H-SS, Park S-CC, Chun J. (2014). Towards a taxonomic coherence between average

531 nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes.
532 *Int J Syst Evol Microbiol* **64**: 346–351.

533 Kirkegaard RH, McIlroy SJ, Kristensen JM, Nierychlo M, Karst SM, Dueholm MS, *et al.* (2017).
534 The impact of immigration on microbial community composition in full-scale anaerobic digesters.
535 *Sci Rep* **7**. e-pub ahead of print, doi: 10.1038/s41598-017-09303-0.

536 Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, *et al.* (2013). Evaluation of
537 general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based
538 diversity studies. *Nucleic Acids Res* **41**: 1–11.

539 Kong Y, Nielsen JL, Nielsen PH. (2005). Identity and Ecophysiology of Uncultured Actinobacterial
540 Polyphosphate-Accumulating Organisms in Full-Scale Enhanced Biological Phosphorus Removal
541 Plants. *Appl Environ Microbiol* **71**: 4076 LP – 4085.

542 Lane DJ. (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid*
543 *Techniques in Bacterial Systematics*. John Wiley and Sons: Chichester, United Kingdom, pp 115–
544 175.

545 Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar a., *et al.* (2004). ARB: A
546 software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

547 Marques R, Santos J, Nguyen H, Carvalho G, Noronha JP, Nielsen PH, *et al.* (2017). Metabolism
548 and ecological niche of Tetrasphaera and Ca. Accumulibacter in enhanced biological phosphorus
549 removal. *Water Res* **122**: 159–171.

550 Martiny JBH, Jones SE, Lennon JT, Martiny AC. (2015). Microbiomes in light of traits: A
551 phylogenetic perspective. *Science (80-)* **350**: aac9823-18.

552 Mcdonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, *et al.* (2018).
553 American Gut : an Open Platform for Citizen Science. *mSystems* **3**: 1–28.

554 McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SMSMSM, *et al.*
555 (2017). MiDAS 2.0: An ecosystem-specific taxonomy and online database for the organisms of
556 wastewater treatment systems expanded for anaerobic digester groups. *Database* **2017**: 1–9.

557 McIlroy SJSJ, Karst SMSMSM, Nierychlo M, Dueholm MSMS, Albertsen M, Kirkegaard RHRH,
558 *et al.* (2016). Genomic and in situ investigations of the novel uncultured Chloroflexi associated with
559 0092 morphotype filamentous bulking in activated sludge. *ISME J* **(In press)**: 1–12.

560 Mikaelyan A, Köhler T, Lampert N, Rohland J, Boga H, Meuser K, *et al.* (2015). Classifying the
561 bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database
562 (DictDb). *Syst Appl Microbiol* **38**: 472–482.

563 Muyzer G, de Waal EC, Uitterlinden AG. (1993). Profiling of complex microbial populations by
564 denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes
565 coding for 16S rRNA. *AEM* **59**: 695–700.

566 Newton IL, Roeselers G. (2012). The effect of training set on the classification of honey bee gut
567 microbiota using the Naïve Bayesian Classifier. *BMC Microbiol* **12**: 221.

568 Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A Guide to the Natural History
569 of Freshwater Lake Bacteria.

570 Nguyen HTT, Le VQ, Hansen AA, Nielsen JL, Nielsen PH. (2011). High diversity and abundance
571 of putative polyphosphate-accumulating Tetrasphaera-related bacteria in activated sludge systems.
572 *FEMS Microbiol Ecol* **76**: 256–267.

573 Parada AE, Needham DM, Fuhrman JA. (2016). Every base matters: Assessing small subunit rRNA
574 primers for marine microbiomes with mock communities, time series and global field samples.
575 *Environ Microbiol* **18**: 1403–1414.

576 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, *et al.* (2018). A
577 standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.
578 *Nat Biotechnol* **36**: 996.

579 Peterson J, Bonazzi V, Starke-Reed P, Read J, Zakhari S, Baker CC, *et al.* (2009). The NIH Human
580 Microbiome Project. *Genome Res* **19**: 2317–2323.

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

583 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal
584 RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* **41**:
585 D590-6.

586 R Core Team. (2016). R: A language and environment for statistical computing. http://www.r-
587 project.org/.

588 Ritari J, Salojärvi J, Lahti L, de Vos WM. (2015). Improved taxonomic assignment of human
589 intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* **16**: 1056.

590 Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD. (2018). TaxAss: Leveraging a Custom
591 Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* **3**: 1–14.

592 RStudio Team. (2015). RStudio: Integrated Development Environment for R.
593 http://www.rstudio.com/.

Seedorf H, Kittelmann S, Henderson G, Janssen PH. (2014). RIM-DB: a taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. *PeerJ* **2**: e494.

Segota I, Long T. (2019). A high-resolution pipeline for 16S-sequencing identifies bacterial strains in human microbiome.

Single- DRR, Sequence NC. (2017). Deblur Rapidly Resolves Single-. **2**: 1–7.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, *et al.* (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**: 457–463.

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

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

Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. (2011). PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**: 1159–1161.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

Wickham H. (2009). ggplot2 - Elegant Graphics for Data Analysis. Springer Science & Business Media.

614 Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, *et al.* (2019). Global diversity and biogeography
615 of bacterial communities in wastewater treatment plants. *Nat Microbiol*. e-pub ahead of print, doi:
616 10.1038/s41564-019-0426-5.

617 Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, *et al.* (2014). Uniting the
618 classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat*
619 *Rev Microbiol* **12**: 635–645.

620 Yilmaz LS, Parnerkar S, Noguera DR. (2011). MathFISH, a web tool that uses thermodynamics-
621 based mathematical models for *in silico* evaluation of oligonucleotide probes for fluorescence *in*
622 *situ* hybridization. *Appl Environ Microbiol* **77**: 1118–1122.

623

Tables:

Table 1: Numbers and percentage of full-length ESVs estimated to belong to novel taxa. ESVs were mapped to SILVA_132_SSURef_Nr99 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).

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	Archaea	5	0	0	0	0	0	0
Activated sludge	RNA	Bacteria	460	35 (7.61%)	1 (0.22%)	0	0	0	0
Activated sludge	DNA	Bacteria	6225	1530 (24.6%)	328 (5.27%)	10 (0.16%)	2 (0.03%)	1 (0.02%)	1 (0.02%)
Anaerobic digester	RNA	Archaea	74	0	0	0	0	0	0
Anaerobic digester	RNA	Bacteria	235	33 (14.0%)	6 (2.55%)	0	0	0	0
Anaerobic digester	DNA	Bacteria	4173	937 (22.5%)	222 (5.32%)	4 (0.096%)	1 (0.02%)	0	0
Combined	Combined	Archaea	75	0	0	0	0	0	0
Combined	Combined	Bacteria	9446	2434 (25.8%)	545 (5.77%)	15 (0.16%)	3 (0.03%)	1 (0.01%)	1 (0.01%)

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

Environment	Library	Kingdom	Species	Genus	Family	Order	Class	Phylum
Activated sludge	RNA	Archaea	1 (50%)	0	0	0	0	0
Activated sludge	RNA	Bacteria	189 (88.7%)	47 (40.9%)	16 (22.5%)	5 (10.2%)	0	0
Activated sludge	DNA	Bacteria	2709 (92.6%)	893 (71.2%)	180 (44.4%)	45 (24.9%)	9 (12.3%	1 (2.94%)
Anaerobic digester	RNA	Archaea	10 (58.8%)	0	0	0	0	0
Anaerobic digester	RNA	Bacteria	117 (91.4%)	50 (51.6%)	21 (30.4%)	6 (12.0%)	3 (9.68%)	0
Anaerobic digester	DNA	Bacteria	1760 (92.5%)	595 (66.0%)	132 (43.9%)	36 (23.1%)	8 (11.43%)	0
Combined	Combined	Archaea	10 (58.8%)	0	0	0	0	0
Combined	Combined	Bacteria	3879 (93.3%)	1284 (72.4%)	260 (50%)	67 (29.9%)	13 (14.4%)	1 (2.44%)

Figures:

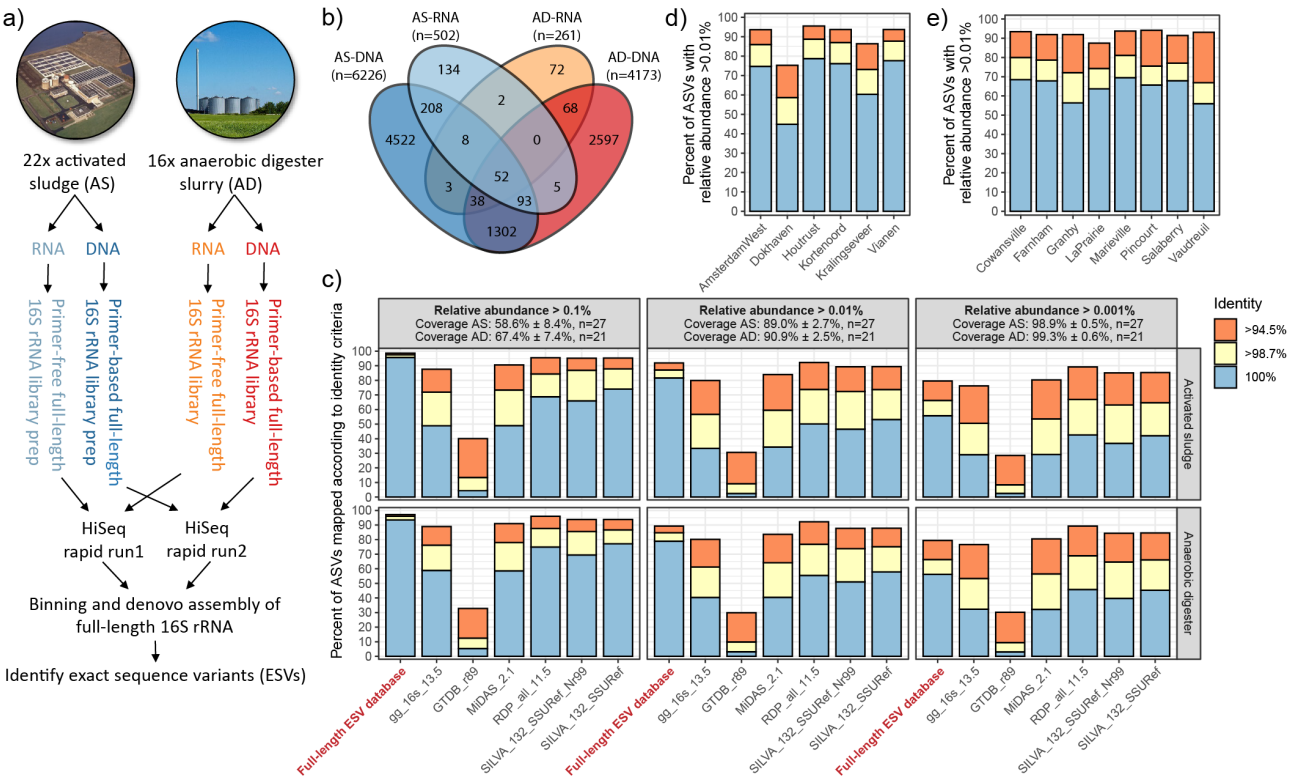

Figure 1. Construction and evaluation of the full-length ESV reference database. a) Preparation of full-length ESVs. Samples were collected from WWTPs and anaerobic digesters, and DNA and RNA were extracted. Purified DNA or RNA were used for preparation of primer-based and "primer-free" full-length 16S rRNA libraries, respectively. These were sequenced and processed bioinformatically to produce a comprehensive ecosystem-specific full-length 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 full-length ESV database and common 16S rRNA reference databases. ASVs were obtained from activated sludge and anaerobic digester samples and filtered based on their relative abundance, before the analysis to uncover the depth of the full-length ESV database. The fraction of the microbial community represented by the remaining ASVs after the filtering (coverage) is shown as mean ± standard deviation across plants. d) Mapping of ASVs from Dutch WWTPs based on raw data from Gonzalez-Martinez et al. (Gonzalez-Martinez *et al.*, 2016). For details, see Figure S3. e) Mapping of ASVs from Canadian WWTPs, based on raw data from Isazadeh et al. 2016 (Isazadeh *et al.*, 2016). For details, see Figure S4.

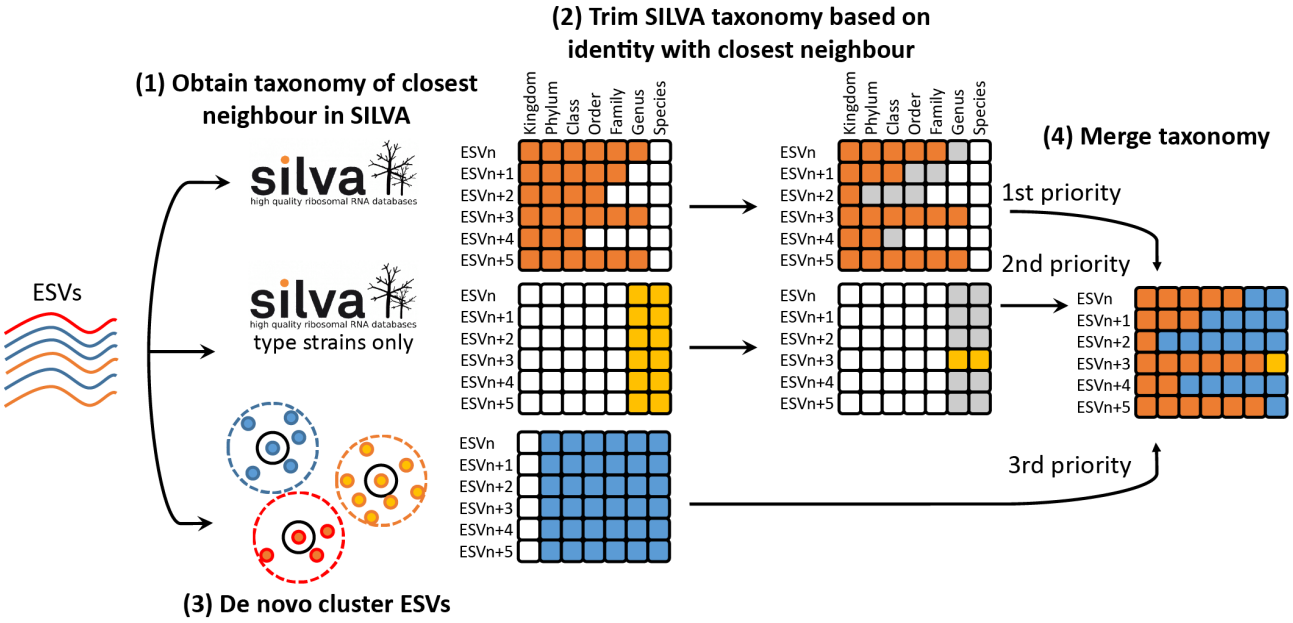

Figure 2. The AutoTax taxonomic framework. (1) Full-length ESVs were first mapped to the SILVA_132_SSURef_Nr99 database to identify the closest relative and the shared percent identity. (2) Taxonomy was adopted from this sequence after trimming, based on percent identity and the taxonomic thresholds proposed by Yarza et al. (Yarza *et al.*, 2014). 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. (Yarza *et al.*, 2014) 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.

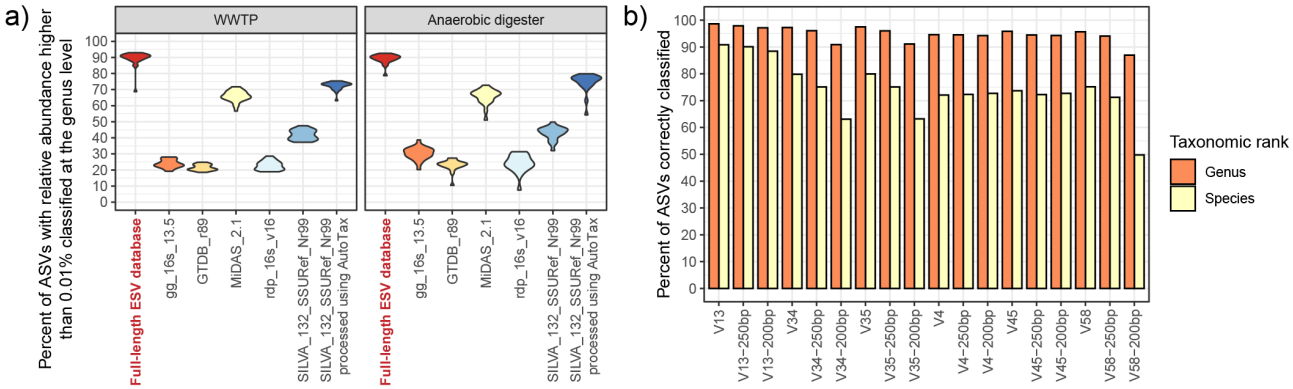

Figure 3. Classification of amplicons. a) Percentage of ASVs from each activated sludge and anaerobic digester sample with a relative coverage of more than 0.01% that were classified to the genus level when classified using the full-length ESV reference database, common reference databases for taxonomic classification, and the SILVA databases processed using AutoTax. b) Classification of *in silico* bacterial 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 the first 250 or 200 bp. V13 (Lane 1991) (Lane, 1991), V34 (Klindworth et al. 2013) (Klindworth *et al.*, 2013), V35 (Peterson et al. 2009) (Peterson *et al.*, 2009), V4 (Apprill et al. 2015) (Apprill *et al.*, 2015), V45 (Parada et al. 2016) (Parada *et al.*, 2016), and V58 (Klindworth et al. 2013) (Klindworth *et al.*, 2013).

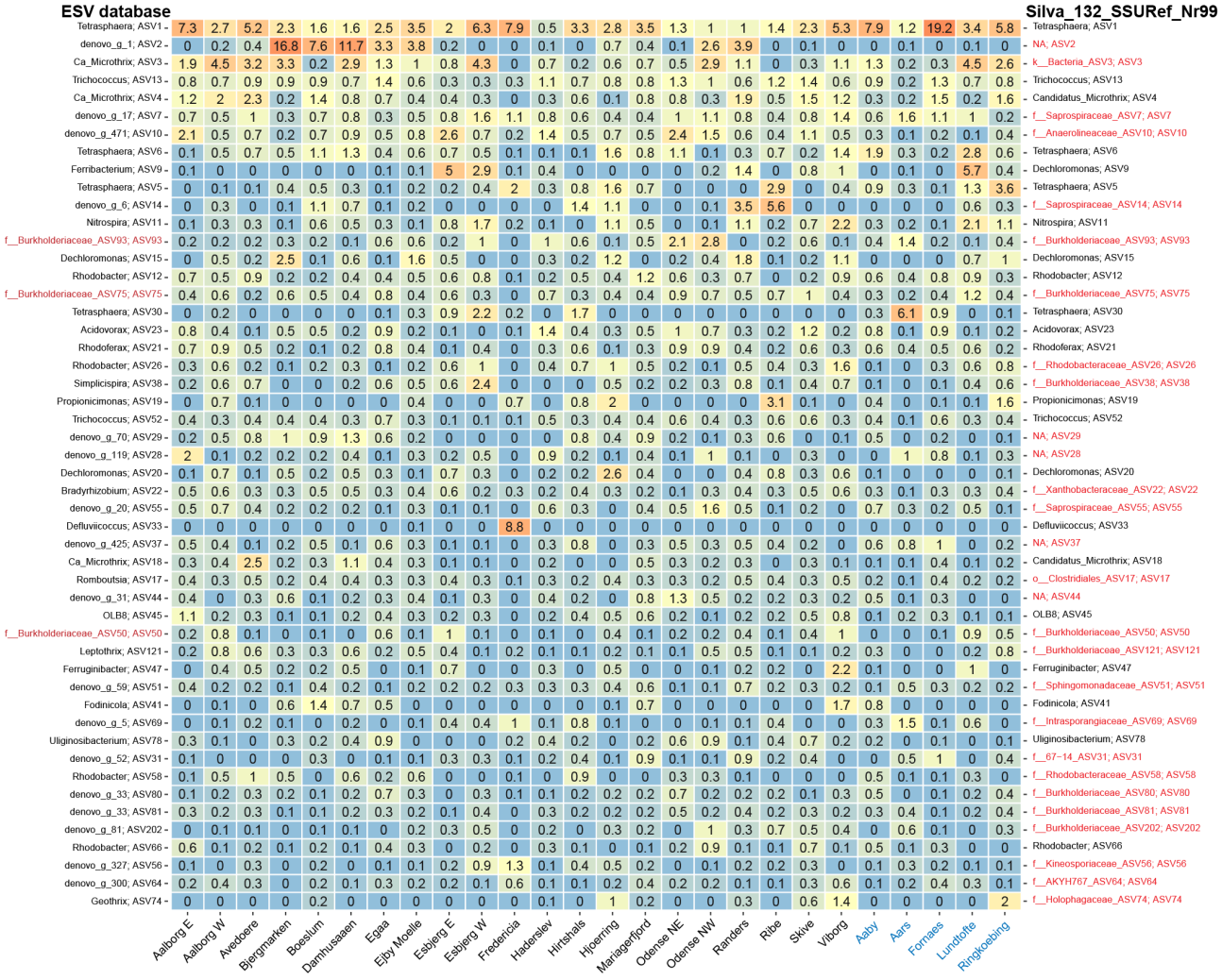

Figure 4. Relative abundance of the top 50 ASVs in the activated sludge samples. ASVs obtained from individual activated sludge samples used to create the full-length ESV database (black labels) and from other plants (blue labels) were classified based on the ESV database as well as the SILVA_132_SSURef_Nr99 database. ASVs not classified at the genus level have been highlighted in red.

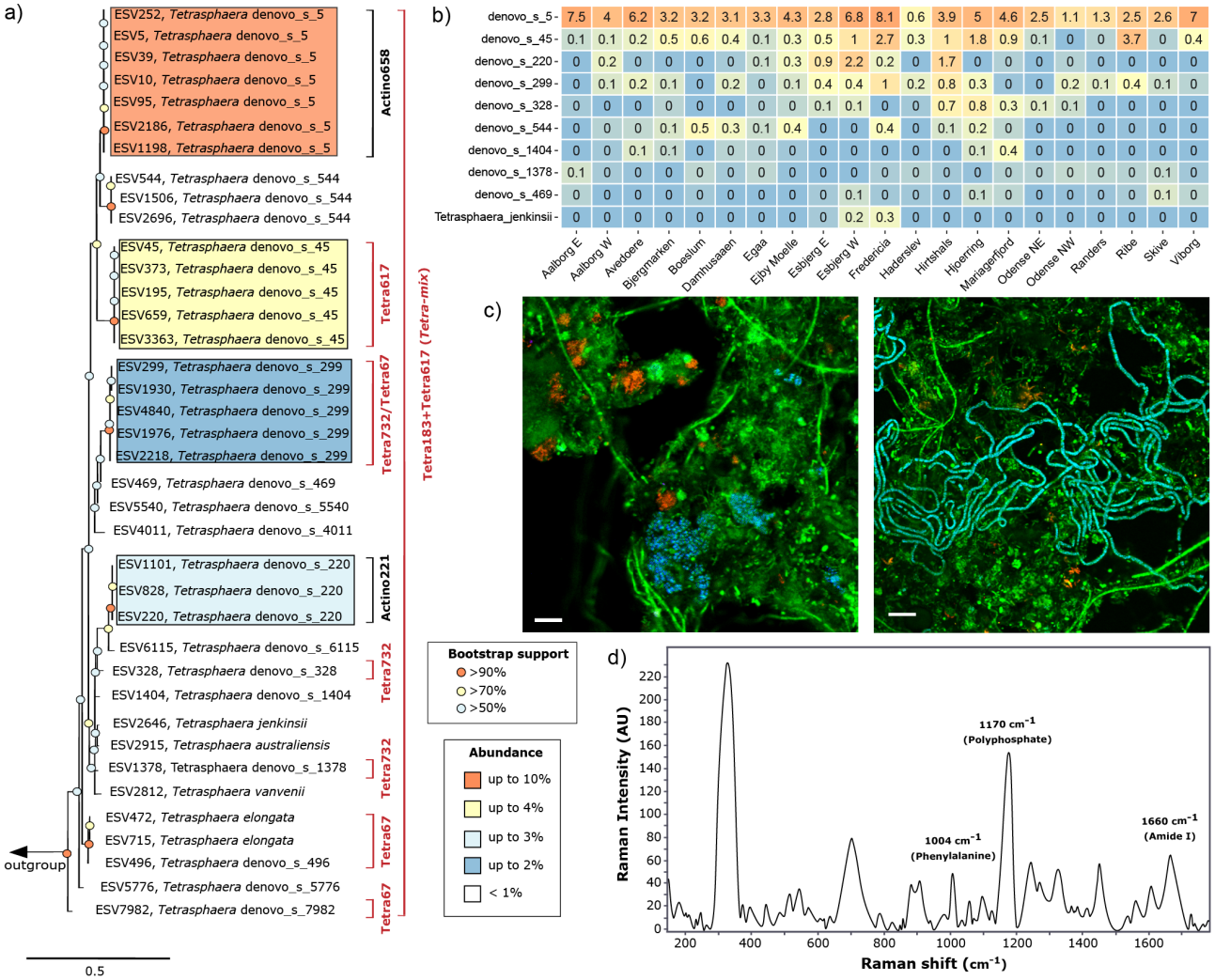

Figure 5. Detailed investigation into the genus *Tetrasphaera*. a) Maximum-likelihood (PhyML) 16S rRNA gene phylogenetic tree of activated sludge *Tetrasphaera* species and coverage of the existing (black) and new FISH probes (red). A 20% conservational filter was applied to the alignment used for the 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% (yellow), and >90% (orange) 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_45 (orange) and tetrads of denovo_s_220 (blue) targeted by the probe Tetra617 and Actino221 (Kong *et al.*, 2005), respectively. Right panel: filaments of denovo_s_328 (cyan) and rods of denovo_s_5 (orange) targeted by the probe Tetra732 and Actino658 (Kong *et al.*, 2005) , 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_45 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.