Generation of comprehensive ecosystems-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax)

Authors: Morten Simonsen Dueholm*[§], Kasper Skytte Andersen[§], Simon Jon McIlroy**, 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: Morten Simonsen Dueholm and 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: md@bio.aau.dk , phn@bio.aau.dk

****Present address:** Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, Qld. 4072, Australia

§ contributed equally to the work

Running title:
Comprehensive ecosystems-specific 16S rRNA gene databases

Abstract

High-throughput 16S rRNA gene amplicon sequencing is an essential 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 absence of a systematic and comprehensive taxonomy for the uncultured majority. Here we demonstrate how high-throughput synthetic long-read sequencing can be applied to create ecosystem-specific full-length 16S rRNA gene amplicon sequence variant (FL-ASV) reference databases that include high-identity references (>98.7% identity) for nearly all abundant bacteria (>0.01% relative abundance) using Danish wastewater treatment systems and anaerobic digesters as an example. In addition, we introduce a novel sequence identity-based approach for automated taxonomy assignment (AutoTax) that provides a complete seven-rank taxonomy for all reference sequences, using the SILVA taxonomy as a backbone, with stable placeholder names for unclassified taxa. The FL-ASVs are perfectly suited for the evaluation of taxonomic resolution and bias associated with primers commonly used for amplicon sequencing, allowing researchers to choose those that are ideal for their ecosystem. The AutoTax taxonomy greatly improves the classification of short-read 16S rRNA gene amplicon sequence variants (ASVs) at the genus- and species-level, compared to the commonly used universal reference databases. Importantly, the placeholder names provide a way to explore the unclassified environmental taxa at different taxonomic ranks, which in combination with *in situ* analyses can be used to uncover their ecological roles.

Introduction:

Microbial communities underpin key biochemical transformations in natural and engineered ecosystems. A deep understanding of these systems requires reliable identification of the microbes present, which can then be associated with their metabolic and ecological 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 and species level (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 (usually 200-500 bp long) is amplified by PCR and sequenced. The amplicons are then clustered into operational taxonomic units (OTUs) or used to infer exact amplicon sequence variants (ASVs) with denoising algorithms such as Deblur (2), DADA2 (3), or Unoise3 (4). The taxonomy of these amplicons is then assigned based on their comparison to a reference database (5–7).

ASVs are often preferred over OTUs because they provide single-nucleotide resolution and can be applied as consistent labels for microbial identification independently of a 16S rRNA gene reference database (8). This approach is used with short-read ASVs in several large-scale projects, including the Earth Microbiome Project (EMP) (9) and the American Gut project (10), 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 the physiology and ecology of key community members, which is crucial if we want to use microbial community structure to predict ecosystem functions or process performance in engineered systems. Firstly, without taxonomic assignment, it is not possible to compare results across studies that have used primers targeting different regions of the 16S rRNA gene. Secondly, short-read ASVs alone do not contain enough evolutionary information to resolve their phylogeny confidently (11, 12). This makes it impossible to report and infer how microbial traits are conserved at different phylogenetic scales. As such, functional properties for uncultured lineages predicted from the annotation of metagenome-assembled genomes (MAGs) with complete rRNA genes (Hiqh-quality MIMAGs standard) (11), or determined through *in situ* studies, cannot be confidently linked to the shorter ASV sequences (12, 13). A robust taxonomic assignment is therefore crucial for cross-study comparisons and the accumulation and dissemination of knowledge relating to uncultured lineages.

74 Taxonomic assignment to ASVs relies on a classifier, e.g., the SINTAX (7) or the RDP
75 classifier (6), which uses statistical algorithms to compare each ASV to a full-length 16S rRNA
76 gene reference database to propose the best estimate of their taxonomy. Confident classification at
77 the lowest taxonomic ranks (genus and species) requires high-identity reference sequences (≥98.7%
78 identity) and a complete seven-rank taxonomy for all references (13). Neither of these criteria is met
79 with the most commonly applied universal reference databases, e.g., Greengenes (14), SILVA (15),
80 and RDP (16), as they lack sequences or taxonomic assignment for many uncultivated
81 environmental taxa. Given the vast diversity predicted for microbial life on Earth (17), it will be
82 some time before reference sequences for all species are generated, and the manual curation of their
83 taxonomy will not be feasible.
84 A potential solution to the problems mentioned above is to create ecosystem-specific
85 reference databases. These can either be ecosystem-curated versions of universal databases or
86 smaller independent databases that only include sequences from the specific ecosystem. The
87 MiDAS 2 database for microbes in biological wastewater treatment systems (18) and the
88 Dictyopteran gut microbiota reference Database (DictDb) (19) are examples of ecosystem-curated
89 versions of the SILVA databases, where the taxonomies for the abundant and process-critical
90 microorganisms are manually curated and maintained. However, the mostly manual taxonomic
91 curation of central reference databases is time-consuming and subjective. The universal reference
92 databases are also clustered at 99% identity or below to keep the unique sequences to a manageable
93 number, which reduces the taxonomic resolution.
94 Examples of independent ecosystem-specific databases include the human intestinal tract 16S
95 taxonomic database (HITdb) (20), the human oral microbiome database (HOMD) (21), the
96 freshwater-specific FreshTrain database (22, 23), the honey bee gut microbiota database (24), and
97 the rumen and intestinal methanogen database (25). While such databases have been shown to
98 improve the rate of classifications for amplicons, they generally contain a relatively limited number
99 of sequences and are therefore associated with an inherent risk of over- or misclassification if the
100 sequence being classified is not represented in the database. To address this issue, Rohwer *et al.*
101 introduced the TaxAss algorithm that classifies amplicons using two reference databases: a
102 universal database and a small ecosystem-specific database (23). Amplicons are first mapped to the
103 ecosystem-specific database to determine the percent identity with the best hit, and those above a
104 user-defined threshold are classified using the ecosystem-specific database. The remaining
105 sequences are classified using the more extensive universal database. While higher rates of

106 classification were achieved, an issue with the approach is the potential for closely related
107 sequences that fall on either side of the user-defined threshold, to receive very different taxonomies,
108 especially if the ecosystem-specific database is not updated to reflect the evolving taxonomy of the
109 universal reference database. As such, while current strategies to create ecosystem-specific
110 databases have shown promise, there are critical issues that need to be resolved before their
111 potential can be realized.

112 The recent development of methods for high-throughput full-length 16S rRNA gene
113 sequencing, e.g., synthetic long-read sequencing on the Illumina platform (26, 27), along with
114 PacBio (28) or Nanopore (29) consensus sequencing, now allows for the generation of millions of
115 high-quality reference sequences within days. Importantly, these technologies now allow for the
116 high-throughput generation of high-identity reference databases with broad coverage of the true
117 diversity. However, improving sequence coverage alone will not solve the problem of poor
118 taxonomic assignments for many uncultured taxa. We have therefore developed the AutoTax
119 pipeline, which provides a simple and efficient strategy for the creation of comprehensive
120 ecosystem-specific taxonomies that cover all seven taxonomic ranks. AutoTax uses the SILVA
121 taxonomy as a backbone and provides stable placeholder names for unclassified taxa, based on *de*
122 *novo* clustering of sequences according to statistically supported identity thresholds for each
123 taxonomic rank (12). Importantly, AutoTax databases are easily updated with subsequent releases
124 of the SILVA taxonomy – avoiding the divergence of generated ecosystem-specific taxonomies
125 with the universal reference database. The strict computational nature of the taxonomy assignment
126 means that it is objective and reproducible. The simplicity of the applied *de novo* clustering also
127 ensures that the placeholder names are maintained even though the database is expanded with
128 additional reference sequences.

129 We demonstrate the potential of the AutoTax method by sequencing almost a million full-
130 length 16S rRNA gene sequences from Danish biological wastewater treatment and bioenergy
131 systems. The sequences were denoised to resolve full-length exact amplicon sequence variants (FL-
132 ASVs) with single-nucleotide resolution. Taxonomy was assigned to the FL-ASVs using the
133 AutoTax pipeline to create an ecosystem-specific reference database. As evidence supporting the
134 value of our approach, mapping of short-read amplicon data revealed that a substantially higher
135 proportion of sequences were matched to high-identity references, and received species and genus
136 level classification when the FL-ASV database was used compared to the much larger public
137 universal reference databases.

138 ## Materials and methods:

139 **General molecular methods**

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

147

148 **Samples and nucleic acid purification**

149 Activated sludge and anaerobic digester biomass were obtained as frozen aliquots (-80°C) from the
150 MiDAS collection (18). Sample metadata is provided in **Table S2**. Total nucleic acids were purified
151 from 500 µL of sample thawed on ice using the PowerMicrobiome RNA isolation kit (MO BIO
152 Laboratories) with the optional phenol-based lysis or with the RiboPure RNA purification kit for
153 bacteria (Thermo Fisher Scientific). Purification was carried out according to the manufacturers'
154 recommendations, except that cell lysis was performed in a FastPrep-24 instrument for 4x 40 s at
155 6.0 m/s to increase the yield of nucleic acids from bacteria with sturdy cell walls (30). The samples
156 were incubated on ice for 2 min between each bead beating to minimize heating due to friction.
157 DNA-free total RNA was obtained by treating 41 µL of the purified nucleic acid with the DNase
158 Max kit (MO BIO Laboratories), followed by clean up using 1.0x RNAClean XP beads with elution
159 into 25 µL nuclease-free water.

160

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

162 Purified RNA obtained from biomass samples was pooled separately for each sample source type
163 (activated sludge or anaerobic digester) to give equimolar amounts of 16S rRNA determined based
164 on peak area in the TapeStation analysis software A.02.02 (SR1). Full-length SSU sequencing
165 libraries were then prepared, as previously described (26). The SSU_rRNA_RT2 (activated sludge)
166 and SSU_rRNA_RT3 (anaerobic digester) reverse transcription primers and the SSU_rRNA_1
167 adaptor were used for the molecular tagging (**Table S1**), and approximately 1,000,000 tagged
168 molecules from each pooled sample were used to create the clonal library. The final library was
169 sequenced on a HiSeq2500 using on-board clustering and rapid run mode with a HiSeq PE Rapid

170 Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2, 265 cycles (Illumina), as previously
171 described (26). Raw sequence reads were binned based on unique molecular tags, *de novo*
172 assembled into synthetic long-read sequences, and trimmed equivalent to *E. coli* position 8 and
173 1507 using the fSSU-pipeline-RNA_v1.2.sh scripts script
174 (https://github.com/KasperSkytte/AutoTax) (26).
175
176 **Primer-based full-length 16S rRNA gene library preparation and sequencing**
177 The purified nucleic acids obtained from the biomass samples were pooled separately for each
178 sample source type (activated sludge or anaerobic digester) with an equal weight of DNA from each
179 sample. Full-length 16S rRNA sequencing libraries were then prepared, as previously described
180 (26). The f16S_rDNA_pcr1_fw1 (activated sludge) or f16S_rDNA_pcr1_fw2 (anaerobic digester)
181 and the f16S_rDNA_pcr1_rv (**Table S1**) were used for the molecular tagging, and approximately
182 1,000,000 tagged molecules from each pooled sample were used to create the clonal library. The
183 final library was sequenced on a HiSeq2500 using on-board clustering and rapid run mode with a
184 HiSeq PE Rapid Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2, 265 cycles (Illumina) as
185 previously described (26). Raw sequence reads were binned based on unique molecular tags, *de*
186 *novo* assembled into synthetic long-read sequences, and trimmed equivalent to *E. coli* position 28
187 and 1491 using the fSSU-pipeline-DNA_v1.2.sh scripts script
188 (https://github.com/KasperSkytte/AutoTax) (26).
189
190 **Extraction of high-quality "full-length" 16S rRNA gene sequences from SILVA**
191 High-quality bacterial and archaeal 16S rRNA gene sequences in the SILVA 138 SSURef NR99
192 ARB-database was selected using the query pintail_slv = 100 and tax_slv = Bacteria* or tax_slv =
193 Archaea*. Bacterial and Archaeal sequences were exported separately in the "fastawide" format
194 after terminal trimming. Bacterial sequences were trimmed between the 27F and 1391R (31) primer
195 binding sites equivalent to position 1,044 and 41,788 in the global SILVA alignment. Archaeal
196 sequences were trimmed between the 20F (32) and the SSU1000ArR (33) primer binding sites
197 equivalent to position 1,041 and 32,818 in the global SILVA alignment. A list of names for "full-
198 length" sequences spanning the positions above was created using the Extract_full-
199 length_16S_rRNA_names_from_SILVA.sh script, which takes advantage of the fact that ARB uses
200 the period to specify terminal gaps and therefore indicates truncated sequences in the exported

FASTA-files. The names were used to select and export the "full-length" bacterial or archaea sequences without trimming from the SILVA ARB-database.

Generation of reference databases using AutoTax

AutoTax was created as a modular multi-step Linux BASH script that 1) generates FL-ASV and full-length 16S rRNA gene operational taxonomic units clustered at 99% identity (FL-OTU) reference sequences from high-quality, full-length 16S rRNA sequences. 2) Assigns a comprehensive seven-rank taxonomy to all reference sequences based on the SILVA taxonomy with the addition of placeholder names for unclassified taxa defined by *de novo* clustering of sequences using specific identity thresholds for each taxonomic rank (12). 3) Produces formatted reference databases, which can be directly used for classification using SINTAX or classifiers in the QIIME 2 framework.

AutoTax combines several software tools (GNU parallel v. 20161222 (34), USEARCH v. 11.0.667 (35), SINA v.1.6.0 (36), and R v.3.5.0 with the following packages: biostrings (37), doParallel (38), stringr (39), data.table (40), tidyr (41), and dplyr (42)) into a single BASH script that otherwise only requires a single FASTA file with the user provided full-length 16S rRNA gene sequences and the FASTA-formatted SILVA_138_SSURef_NR99_tax_silva reference database as input. The script, as well as a docker container image with all required software (except USEARCH, as the required 64 bit version it is non-free and must be purchased from https://drive5.com/usearch), is available on the GitHub repository https://github.com/KasperSkytte/AutoTax. The script is composed of separate, individual BASH functions to both allow for customization of the script as well as unit testing using the BASH Automated Testing System (https://github.com/bats-core/bats-core) where possible. Core functions of AutoTax are briefly described below. Expanded descriptions can be found in the supplementary information (**Text S1**).

Resolving full-length 16S rRNA amplicon sequence variant (FL-ASV)

Input sequences are oriented based on the SILVA 138 SSURef NR99 database using the *usearch -orient* command, dereplicated using *usearch -fastx_uniques* with the *-sizeout, -strand plus*, and *-threads 1* options, and finally denoised to produce the FL-ASVs using the *usearch -unoise3* command with the *-minsize 2* option.

233 *Preparation of chimera-filtered full-length 16S rRNA OTUs (FL-OTUs)*
234 Dereplicated sequences (before denoising) from above are clustered at 99% sequence identity using
235 *usearch -cluster_smallmem* command with the *-id 0.99*, *-maxrejects 0, -centroids*, and *-sortedby*
236 *size* options. Potential chimeras are identified and extracted using the *usearch -uchime2_ref*
237 command with the *-strand plus, -mode sensitive*, and *-chimeras* options with the FL-ASVs from
238 above as the reference database. The chimeras are finally removed to create the final FL-OTUs
239 using the *usearch -search_exact* command with the *-strand plus* and *-dbnotmatched* options.
240
241 *Taxonomy assignment*
242 A complete taxonomy from kingdom to species is automatically assigned to each FL-ASV. In brief,
243 the AutoTax script identifies the closest relative of each FL-ASV in the SILVA database using the
244 *usearch -usearch_global* command, obtains the taxonomy for this sequence, and discards
245 information at taxonomic ranks not supported by the sequence identity and the thresholds for
246 taxonomic ranks proposed by Yarza *et al*. (12). The identity thresholds used for each of the
247 taxonomic ranks are: Phylum (75.0%), Class (78.5%), Order (82.0%), Family (86.5%), Genus
248 (94.5%), and Species (98.7%). For the species-level classification, the script identifies all type
249 strains within the species-level threshold in the SILVA database and assigns a species-level
250 classification to the FL-ASV if only a single species fits within the threshold. In addition, FL-ASVs
251 are *de novo* clustered using the *usearch -cluster_smallmem* command using the thresholds for each
252 taxonomic rank. The *de novo* clusters are labeled according to the format denovo_x_y, where x is a
253 one-letter abbreviation for the taxonomic rank (k, p, c, o, f, g, and s), and y represents the FL-ASV
254 number of the cluster centroid for each taxon. These labels act as a placeholder taxonomy, where
255 the SILVA taxonomy does not provide any taxonomy information.
256
257 **Amplicon sequencing and analyses**
258 Bacterial community analyses were performed by amplicon sequencing of the V1-3 variable region
259 of the 16S rRNA gene as previously described (43) using the 27F (5'-
260 AGAGTTTGATCCTGGCTCAG-3') (31) and 534R (5'-ATTACCGCGGCTGCTGG-3') (44)
261 primers. Forward reads were processed using USEARCH v.11.0.667. Raw fastq files were filtered
262 for phiX sequences using *usearch -filter_phix*, trimmed to 250 bp using *usearch -fastx_truncate -*
263 *trunclen 250*, and quality filtered using *usearch -fastq_filter* with the *-fastq_maxee 1.0* option. The
264 sequences were dereplicated using *usearch -fastx_uniques* with the *-sizeout* option. Exact amplicon

265 sequence variants (ASVs) were resolved using *usearch -unoise3* (4). ASV-tables were created by
266 mapping the raw reads to the ASVs using *usearch -otutab* with the *-zotus* and *-strand both* options.
267 Taxonomy was assigned to ASVs using *usearch -sintax* with *-strand both* and *-sintax_cutoff 0.8*
268 (13).
269
270 **Construction of phylogenetic trees and primer evaluation**
271 FL-ASVs aligned to SILVA 138 NR99 ARB-database was obtained from the AutoTax output
272 (temp/FL-ASVs_SILVA_aligned.fa) and loaded into SILVA 138 NR99 ARB-database. All
273 bacterial FL-ASVs were selected and exported as a FASTA-file using the ssuref:bacteria positional
274 variability by parsimony filter. A rough tree was created from the alignment using FastTree v.
275 2.1.10 (45) with the *-nt -gtr -gamma* options and loaded into ARB. The specificity of commonly
276 used amplicon primers was determined for each FL-ASV using the *analyze_primers.py* script from
277 Primer Prospector v. 1.0.1 (46). The specificity of primer sets was defined based on the overall
278 weighted scores (OWS) for the primer with the highest score as follows: Perfect hit (OWS = 0),
279 partial hit (OWS > 0, and ≤ 1), poor hit (OWS > 1). A comma-separated table with the specificity of
280 each primer set for each FL-ASV was made in R, loaded into ARB, and used to color the tree.
281
282 **Data analyses and visualization**
283 USEARCH v.11.0.667 was used for mapping sequences to references with *-usearch_global -id 0 -*
284 *maxrejects 0 -maxaccepts 0 -top_hit_only -strand plus,* unless otherwise stated. Data were imported
285 into R v.3.6.3 (47) using RStudio IDE (48), aggregated using the tidyverse package v.1.2.1
286 (https://www.tidyverse.org/), and analyzed and visualized using ggplot2 v.3.1.0 (49) and Ampvis2
287 v.2.4.0 (50).
288
289 **Data and code availability**
290 Raw and assembled sequencing data is available at the European Nucleotide Archive
291 (https://www.ebi.ac.uk/ena) under the project number PRJEB26558. The AutoTax script is
292 available at https://github.com/KasperSkytte/AutoTax. AutoTax processed FL-ASV and FL-OTU
293 expanded reference database in SINTAX and QIIME format, is available at
294 https://figshare.com/articles/Data_used_in_AutoTax_paper/12377741/1. The AutoTax processed
295 SILVA 138 SSURef NR99 database in SINTAX, and QIIME format is available at

https://doi.org/10.6084/m9.figshare.12366626. R-markdown scripts used for data analyses and figures are available at https://github.com/msdueholm/Publications/tree/master/Dueholm2019a.

298 ## Results and Discussion:

299 **Sampling and high-throughput sequencing of full-length 16S rRNA sequences**

300 To obtain 16S rRNA gene reference sequences for Danish wastewater treatment plants (WWTPs)
301 and anaerobic digesters (ADs), we sampled biomass from 22 typical WWTPs and 16 ADs treating
302 waste activated sludge located at Danish wastewater treatment facilities (**Table S2**). These facilities
303 represent an important engineered ecosystem containing complex microbial communities of both
304 bacteria and archaea, with the vast majority of microbes being uncultured and poorly characterized
305 (51).

306 DNA and RNA were extracted and used for synthetic long-read 16S rRNA gene sequencing
307 using both a primer-based and primer-free approach (26) (**Figure 1a**). A total of 926,507 full-length
308 16S rRNA gene sequences were obtained after quality filtering. These were denoised with
309 UNOISE3 to generate a comprehensive reference database of 9,521 FL-ASVs with an error-rate
310 below the detection limit according to theoretical calculations and analysis of 7,816 sequences
311 previously obtained from the eight strain ZymoBIOMICS Microbial Community DNA Standard
312 (26) (**Text S1**).

313 To estimate the number of FL-ASVs belonging to novel taxa, FL-ASVs were mapped to the
314 SILVA 138 SSURef NR99 database (15) using global mapping with USEARCH, and the identity of
315 their closest relatives was compared to the thresholds for taxonomic ranks proposed by Yarza *et al.*
316 2014) (**Table 1**). The majority of the FL-ASVs (94%) had references in the SILVA database above
317 the genus-level threshold (identity >94.5%), but 26% lacked references above the species-level
318 threshold (identity > 98.7%), which are crucial for confident taxonomy assignment to ASVs.

319

320 **The FL-ASVs have better ecosystem coverage than universal reference databases**

321 To evaluate if the FL-ASV database contained high-identity references for all bacteria in the
322 ecosystem, we mapped V1-3 ASVs obtained from two sources: (1) the same samples used to create
323 the FL-ASVs, and (2) samples from unrelated Danish WWTP and ADs. The ecosystem-specific
324 FL-ASV database (9,521 seq.) included more high-identity references (>98.7% identity) for the
325 abundant ASVs (relative abundance cutoff at 0.01%) in all samples analyzed, compared to the much
326 larger universal databases: MiDAS 2.1 (548,447 seq.) (18), SILVA 138 SSURef NR99 (510,984
327 seq.) (15), SILVA 138 SSURef (2,225,272 seq.) (15), GreenGenes 16S v.13.5 (1,262,986 seq.) (14),
328 and the full RDP v.11.5 (3,356,808 seq.) (16) (**Figure 1b and Figure S1**). ASVs were also mapped
329 to the 16S rRNA database derived from the Genome Taxonomy Database (GTDB) release 89

330 (17,460 seq.) (52). However, this database lacked high-identity references for almost all ASVs. The
331 poor coverage likely relates to the fact that 16S rRNA genes often fail to assemble in MAGs
332 produced by short-read sequencing data (53). This problem will likely disappear in the future with
333 the introduction of more high-quality MAGs with complete rRNA genes into the GTDB as a result
334 of long-read sequencing technologies such as Nanopore and PacBio (54, 55).
335 When the rare biosphere was included in the analysis (relative abundance cutoff at 0.001%),
336 a decrease in the percentage of ASVs with high-identity reference sequences was observed (**Figure**
337 **1b**). This may be a problem in ecosystems with high diversity, such as soil and sediments (9),
338 where low-abundant microbes constitute a considerable fraction of the community, but also in
339 engineered systems where transient or low abundant bacteria, such as pathogens or bacteria
340 degrading micropollutants, may be important. To get a better representation of the rare biosphere,
341 we created an additional reference database, which besides the FL-ASVs, contained the chimera-
342 filtered, full-length 16S rRNA gene sequences clustered at 99% identity (FL-OTUs). This database
343 greatly increased the coverage for the rare biosphere because it includes FL-OTUs that represent
344 singletons. However, the improved coverage is achieved at the expense of taxonomic resolution
345 (see later).
346 Since only Danish WWTPs and ADs were used to establish the comprehensive high-identity
347 FL-ASV reference database, published amplicon data from non-Danish WWTPs (56, 57) was also
348 evaluated (**Figure 1c-d, and S2**). Compared to the analyzed universal reference databases, the
349 Danish reference FL-ASVs performed better or as well for most of the investigated non-Danish
350 WWTPs, which indicates that even with less than 10,000 sequences, the database covers many of
351 the microbes that are common to WWTP across the world, particularly for systems with nutrient
352 removal. We anticipate that our ongoing sampling of more than 1,000 WWTPs and AD systems
353 across all seven continents and different process-configurations (MiDAS Global,
354 https://www.midasfieldguide.org/global) for high-throughput full-length 16S rRNA gene
355 sequencing will provide references for the region-specific taxa in the near future, providing a
356 comprehensive database of reference sequences for this ecosystem.
357
358 **A new comprehensive taxonomic framework**
359 A major limitation in the classification of amplicon data from environmental samples is the lack of
360 genus and species names for many uncultivated bacteria in the universal reference databases. To
361 address this, we developed a simple taxonomic framework (AutoTax), which provides a consistent

362 taxonomic classification of all reference sequences to all seven taxonomic ranks using identity
363 thresholds (**Figure 2**).
364 In the AutoTax pipeline, the FL-ASVs (or FL-OTUs) are first mapped to the SILVA 138
365 SSURef NR99 database, which provides the taxonomy and percent identity of the closest relative in
366 the database. This taxonomy is assigned to the FL-ASV down to the taxonomic rank supported by
367 the sequence identity thresholds proposed by Yarza *et al.* (2014) (**Table 1**). Because of the stringent
368 mapping and the taxonomy trimming based on percent identity, we obtained an overall better
369 taxonomy assignment compared to commonly used classifiers, as revealed by a leave-one-out
370 classification test (**Figure S3**).
371 Since species-level classification is desired wherever possible and the official SILVA
372 taxonomy for bacteria and archaea is not curated at the species level (58), FL-ASVs were also
373 mapped to the 16S rRNA gene sequences from type strains extracted from the SILVA 138 SSURef
374 NR99 database as these carry official species names. Species-level classifications were assigned to
375 the FL-ASVs if they shared more than 98.7% identity with only one species. If the FL-ASVs
376 matched more than one, they were not classified at the species level due to the high risk of
377 misclassification.
378 Because the SILVA taxonomy does not provide a complete seven-rank taxonomy for all
379 sequences, the missing classifications are covered with a *de novo* placeholder taxonomy. This
380 taxonomy was created based on the clustering of the FL-ASVs at identity thresholds corresponding
381 to each taxonomic rank (12). The clusters were labeled according to the format denovo_x_y, where
382 x is a one-letter abbreviation for the taxonomic rank (k, p, c, o, f, g, and s), and y represents the
383 number of the FL-ASV which is the cluster centroid of the particular taxon. Because the applied
384 clustering algorithm processes the sequences sequentially in the order they appear in the input file,
385 sequences are always clustered in the same way, even if additional FL-ASVs are later added to the
386 database. This strategy may not always yield the most optimal clusters, but the reproducibility is
387 critical if the clusters are to be used as a robust placeholder taxonomy.
388 Merging of the SILVA- and the *de novo*-based taxonomies resulted in a few conflicts, e.g.,
389 where different FL-ASVs from the same species associate with more than one genus. In such cases,
390 the genus-level classification for the centroid FL-ASV is adapted for all FL-ASVs within that
391 species. However, these types of conflicts only applied to lower rank taxa (species), which were
392 located close to the taxonomic threshold of the higher rank taxa (genus), and it only affected the
393 classification of a low number of FL-ASVs (approx. 1%).

394 As AutoTax is based on the SILVA taxonomy, the taxonomy generated will change if
395 another version of the SILVA SSURef NR99 reference database is used. Accordingly, users must
396 specify which version of the SILVA database has been used when they publish databases created
397 with AutoTax. We recommend that AutoTax generated databases are updated when there is a new
398 version of the SILVA database. This ensures that the taxonomy is in agreement with the current
399 central taxonomy.
400
401 **Taxonomy assignment to FL-ASVs with AutoTax**
402 AutoTax provided placeholder names for many previously undescribed taxa in our FL-ASV
403 database (**Table 2, Figure S4**). 95% of all species, 73% of all genera, 47% of all families, and 24%
404 of all orders obtained placeholder names from the *de novo* taxonomy and would otherwise have
405 remained unclassified. The novel taxa were affiliated with several phyla, especially the
406 Proteobacteria, Planctomycetota, Patescibacteria, Firmicutes, Chloroflexi, Bacteroidota,
407 Actinobacteriota, and Acidobacteriota (**Figure S4**). A prominent example is the Chloroflexi, where
408 5 out of 14 orders, 26 out of 34 families, and 142 out of 152 genera observed were assigned a *de*
409 *novo* placeholder taxonomy. We believe that this will have important implications for future studies
410 and the accumulation and transfer of knowledge about these taxa in WWTPs, given their high
411 diversity and abundance, and their association with serious operational problems related to the
412 settling of activated sludge (bulking) and foaming (59, 60). It should be noted that the placeholder
413 taxonomy does not provide the same degree of support as traditional phylogenetic analyses, and
414 should only be used until an official taxonomy is established.
415
416 **Improved classification of ASVs from WWTP and anaerobic digesters**
417 To benchmark the FL-ASV database, we classified V1-3 amplicon data obtained from activated
418 sludge and anaerobic digester samples (**Table S2**) using this database with the SINTAX-classifier
419 and compared the results to classifications obtained using the universal reference databases (**Figure**
420 **3a**). With the use of the FL-ASV database, many more of the abundant ASVs (relative
421 abundance >0.01%) were classified to the genus and species level (89.9±4.3% and 78.5±4.0%,
422 respectively), compared to the use of popular public reference databases, including SILVA 138
423 SSURef NR99 (30.4±3.5% and 0%), GreenGenes 16s v. 13.5 (24.5±4.4% and 1.4±0.4%), GTDB
424 r89 (22.1±2.8% and 11.6±1.2%), the RDP 16S v16 training set (20.3±4.0% and 0%), and even the

425 MiDAS 2.1 (59.7±4.5% and 0.4±0.3%) - which is a manually ecosystem-specific curated version of
426 the SILVA 123 SSURef NR99 database.
427 We have here shown that increased coverage of the rare biosphere can be achieved by
428 adding FL-OTUs clustered at 99% to the FL-ASV database. The FL-OTU expanded database
429 increased the classification rate at the genus level for the abundant bacteria (relative abundance >
430 0.01%) (94.2±1.4% vs. 89.9±4.3%), but reduced classification at the species level (67.7±2.6% vs.
431 78.5±4.0%) (**Figure 3a**). When ASVs from the rare biosphere (relative abundance > 0.001%) were
432 included in the analysis, the advantage of including FL-OTUs for genus-level classification was
433 even more pronounced (93.0±0.6% vs. 80.7±1.7%), and an improved classification rate at the
434 species level was also observed (69.9±1.5% vs. 66.8±2.4%) (**Figure S5**). We hypothesize that the
435 reduced classification rate at the species level for the abundant ASVs is caused by sequencing errors
436 as well as low-divergence chimera, which cannot be detected by chimera filtering (61). It should be
437 noted that FL-OTUs are problematic for databases that are to be maintained and updated with
438 additional references in the future. This is because the placeholder taxonomy will change if FL-
439 OTUs are removed, and sequencing errors and chimeras are propagated if they are kept. We
440 therefore recommend that FL-OTUs are only added for exploratory purposes.
441 To investigate the influence of AutoTax on taxonomic assignment, independent of our
442 ecosystem-specific database, we applied the pipeline directly to high-quality, full-length sequences
443 from the SILVA 138 SSURef_NR99 database. This increased the percentage of ASVs classified
444 with SILVA at the genus level from 30.4±3.5% to 63.7±4.8%, and at the species-level from 0% to
445 28.2±2.4%, suggesting that the large universal databases would also benefit from the use of
446 AutoTax. An advantage of the AutoTax processed SILVA database, which we have made publicly
447 available, is that the placeholder taxonomy is universally applicable, providing a unique opportunity
448 for studying the ecology of unclassified taxa across ecosystems.
449
450 **Taxonomic resolution of ASVs in combination with a comprehensive reference database**
451 Classification of amplicon sequences can be challenging due to the limited taxonomic information
452 in short-read sequences (12, 13). However, this may change with the access to reference databases
453 with perfect references for the majority of all ASVs and a complete seven-rank taxonomy for all
454 reference sequences. To determine the confidence of the amplicon classification in this scenario, we
455 extracted ASVs *in silico* from the bacterial FL-ASVs corresponding to commonly amplified 16S
456 rRNA regions, including full-length amplicons. These ASVs were classified against the FL-ASV

457 database. We then calculated the fraction of amplicons correctly classified to the same genus and
458 species as their corresponding FL-ASVs (**Figure 3b and Figure S6a**). Nearly all ASVs (95-99%)
459 were assigned to the correct genus and most (72-91%) to the right species, depending on the
460 taxonomic conservation of 16S rRNA region covered by the *in silico* amplicons. The primers
461 targeting the V1-3 variable region performed exceptionally well for species-level identification
462 (90.7% correct classifications), which is the same as for the full-length 16S rRNA gene amplicons.
463 The commonly used primers targeting the V4 variable region were the worst (72.5% correct
464 classifications). Very few of the sequences that did not receive the same taxonomic classification as
465 their source reference sequence where misclassified (<0.2% at genus-level and <0.8% at species-
466 level), with the majority not receiving any classification at the specific taxonomic rank (**Figure**
467 **S6a**).
468 Sequencing costs on the Illumina platforms can be reduced considerably if single reads are
469 used instead of merged reads. To evaluate the effect of reduced amplicon length, we compared the
470 classification of 200 bp forward reads and reverse reads to those of full-length amplicons (**Figure**
471 **3b**). The decrease in the percentage of correct classifications was highly dependent on the 16S
472 rRNA region targeted and from which direction the single reads were obtained. For the ecosystem
473 studied here, the V1-3 and V4-5 forward read provided almost the same specificity as the full-
474 length amplicons, whereas the reverse reads performed much worse for species-level classification.
475 For the V3-4, V5-7, and V5-8, the reverse reads performed better than the forward reads, revealing
476 the importance of choosing the right direction for single read amplicons.
477 To evaluate the effect of sequencing errors and low-divergence chimeras in the reference
478 database, we also classified the *in silico* ASVs against the reference databases expanded with the
479 FL-OTUs (**Figure S6b**). The result confirmed our prior observations that the inclusion of error
480 prone references had a negative impact on our ability to classify short-read amplicons correctly;
481 however, the effect was marginal at the genus-level. Full-length amplicons were less affected
482 (**Figure S6b**), highlighting a clear advantage of using longer amplicons in combination with
483 universal databases, which are likely to contain sequencing errors and low-divergence chimeras
484 despite chimera filtering (61).
485 Overall, the analysis demonstrated that confident classification of short-read ASV sequences
486 at the genus to species level is possible. However, it requires a reference database with a complete
487 seven-rank taxonomy and perfect references for the majority of all ASVs. The scripts made

488 available with this study can be used to confirm whether this is the case for samples from other
489 studies.
490
491 **Ecosystem-specific evaluation of primer-bias**
492 When choosing primers for amplicon sequence analyses, it is essential also to take primer-bias into
493 account as some primer sets may result in ecosystem-relevant species being severely
494 underestimated or absent from the analyses (30). The ecosystem-specific FL-ASV databases
495 provide a near-perfect reference to determine the theoretical coverage of different primer sets for
496 the given ecosystem so that an informed selection can be made (46, 62). It should be noted that
497 primers used to generate the full-length 16S rRNA sequences for FL-ASV databases may introduce
498 a bias, and we have here include primer-free (RNA-based) libraries to account for this. Evaluation
499 of different primer sets using our FL-ASV database revealed a clear taxonomic bias associated with
500 several primer sets (**Figure 3c**). The primers targeting the V4-5 and V5-8 regions had the best
501 coverage of the FL-ASVs. The V5-7 primers demonstrated a very poor coverage. Because the FL-
502 ASVs were trimmed after the forward priming site of the V1-3 primers, we were not able to
503 evaluate the coverage of this primer pair here. However, it has previously been shown that the
504 primers have a good overall agreement with metagenomic data for wastewater treatment systems
505 and capture most of the process-critical organisms (31).
506
507 **Perspectives**
508 High-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment
509 (AutoTax) now allow individual research groups to develop their own FL-ASV ecosystem-specific
510 reference databases for community profiling analyses. In addition, such databases can be used to
511 evaluate the ecosystem-specific coverage and specificity of amplicon primers and fluorescence *in*
512 *situ* hybridization (FISH) probes. The high quality of the FL-ASVs furthermore allows for the
513 design of new primers and probes with improved confidence of the coverage and taxonomic
514 resolution when applied within the target ecosystem. Collectively, the approach importantly allows
515 for the identification and subsequent characterization of novel numerically important taxa for the
516 specific environment, which would have otherwise been overlooked.
517 We acknowledge that the ability to generate new 16S rRNA gene reference databases quickly
518 poses a risk for the development of several competing divergent taxonomies. We, therefore,
519 recommend the custom databases are only used for exploratory purposes and combined with

520 traditional phylogenetic analyses of key taxa. Ecosystem-specific databases that are broadly
521 applied, such as the MiDAS database, should be created as open-source community efforts, and
522 universal reference databases processed with AutoTax should be published and maintained in
523 agreement with current developers of such databases. An important benefit of this approach is that
524 the placeholder taxonomy can be used as a common language within the field, or in the case of
525 universal reference databases, across all fields. This has major implications, e.g., in wastewater
526 treatment systems, as it allows for the identification of unclassified taxa that are process-critical and
527 decisive for process performance. If the current universal reference databases are used, the majority
528 of ASVs will not get a genus-level classification, making it impossible to compare their prevalence
529 across studies. Given that hundreds of amplicon-based studies are carried out every year worldwide,
530 a considerable amount of useful information is lost when the data generated across studies is not
531 comparable.
532 We have chosen to use the SINTAX-classifier for our analyses because it applies a simple
533 algorithm that does not require training, and we expect the results to be less biased. However, some
534 classifiers which use Bayesian inference, e.g., q2-feature-classifier (5) may yield better
535 classifications. Kaehler et al. (63) recently demonstrated what environment-specific taxonomic
536 abundance information could be used as weights for such classifiers to improve the accuracy of the
537 taxonomy assignment. This approach is interesting because the abundance of raw full-length 16S
538 rRNA gene sequences used to generate the FL-ASV database may be used as phylogenetically
539 informative weights for the specific ecosystem.
540 We used the SILVA SSURef NR99 as the backbone taxonomy for AutoTax in this study, as this
541 is currently the most comprehensive database. However, we anticipate the GTDB may replace
542 SILVA in a future release of AutoTax when more high-quality genomes and MAGs with 16S rRNA
543 genes (MIMAG standard (11)) are added to the database as the result of advances in long-read
544 sequencing technology (Nanopore and PacBio). This will importantly link the 16S rRNA gene
545 taxonomy with that derived from the more robust phylogenomic-based analyses, creating a unified
546 language across the field of microbiology.
547
548 **Acknowledgments**
549 We would like to thank the 22 wastewater treatment plants involved in the project for providing
550 samples. This work was supported by the Danish Research Council [6111-00617A to P.H.N.]; and
551 the Villum foundation [13351 to P.H.N.].

552

553 **Conflict of Interest**

554 The authors declare no conflict of interest.

555

556 ## References:

557 1. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: A
558 phylogenetic perspective. Science 350: aac9823-18.

559 2. Amir A, Daniel M, Navas-Molina J, Kopylova E, Morton J, Xu ZZ, Eric K, Thompson L,
560 Hyde E, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single- Nucleotide
561 Community Sequence Patterns. Am Soc Microbiol 2:1–7.

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

564 4. Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon
565 sequencing. bioRxiv 81257.

566 5. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory
567 Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences
568 with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90.

569 6. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid
570 assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol
571 73:5261–5267.

572 7. Edgar R. 2016. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS
573 sequences. bioRxiv 074161.

574 8. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace
575 operational taxonomic units in marker-gene data analysis. ISME J 11:2639–2643.

576 9. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A,
577 Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y,

González A, Morton JT, Mirarab S, Xu ZZ, Jiang L, Haroon MF, Kanbar J, Zhu Q, Song SJ, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Agosto Rivera JL, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll JL, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart HP, Guyton K, Haig SJ, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler A V., Krezalek M, Kueneman J, Lamendella R, Landon EM, Lanede Graaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ, Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson C, Rockne KJ, Rodriguezl SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M, Schrenzel MD, Scott NM, Seger RL, Seguinorlando A, Seldin L, Seyler

610 LM, Shakhsheer B, Sheets GM, Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J,
611 Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling M, Steinberg PD, Stephens B, Stevens
612 MA, Taghavi S, Tai V, Tait K, Tan CL, Taş N, Taylor DL, Thomas T, Timling I, Turner BL,
613 Urich T, Ursell LK, Van Der Lelie D, Van Treuren W, Van Zwieten L, Vargas-Robles D,
614 Thurber RV, Vitaglione P, Walker DA, Walters WA, Wang S, Wang T, Weaver T, Webster
615 NS, Wehrle B, Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR,
616 Whittingham LA, Willerslev E, Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld
617 J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A communal catalogue reveals Earth's multiscale
618 microbial diversity. Nature 551:457–463.

619 10. Mcdonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA,
620 Behsaz B, Brennan C, Chen Y, Goldasich D, Dorrestein PC, Dunn RR, Fahimipour AK,
621 Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C,
622 Jackson MA, Janssen S, Jeste D V, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J,
623 Leach J, Marotz C, Meleshko D, Melnik A V, Metcalf JL, Mohimani H, Montassier E,
624 Rahnavard G, Robbins-pianka A, Sangwan N, Shorenstein J, Smarr L, Vázquez-baeza Y,
625 Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, Gut A, Jt M, Gonzalez A, Ackermann G,
626 Behsaz B, Brennan C, Chen Y, Goldasich L, Pc D, Rr D, Jl G, Hugenholtz P, Humphrey G,
627 Huttenhower C, Ma J, Janssen S, Jiang L, St K, Knights D, Kosciolek T. 2018. American
628 Gut : an Open Platform for Citizen Science. mSystems 3:1–28.

629 11. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz
630 F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A,
631 Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM,
632 Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ,
633 Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen
634 JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I,
635 Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L,
636 Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified
637 genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.
638 Nat Biotechnol 35:725–731.

639 12. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB,
640 Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and

641 uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol
642 12:635–645.

643 13. Edgar RC. 2018. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences.
644 PeerJ 6:e4652.

645 14. Desantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu
646 P, Andersen GL. 2006. Greengenes , a Chimera-Checked 16S rRNA Gene Database and
647 Workbench Compatible with ARB. Appl Environ Microbiol 72:5069–5072.

648 15. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013.
649 The SILVA ribosomal RNA gene database project: Improved data processing and web-based
650 tools. Nucleic Acids Res 41:D590-6.

651 16. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A,
652 Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput
653 rRNA analysis. Nucleic Acids Res 42:D633-42.

654 17. Louca S, Mazel F, Doebeli M, Parfrey LW. 2019. A census-based estimate of earth's
655 bacterial and archaeal diversity. PLoS Biol 17:1–30.

656 18. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM, Albertsen
657 M, Nielsen PH. 2017. MiDAS 2.0: An ecosystem-specific taxonomy and online database for
658 the organisms of wastewater treatment systems expanded for anaerobic digester groups.
659 Database 2017:1–9.

660 19. Mikaelyan A, Köhler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A. 2015.
661 Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic
662 reference database (DictDb). Syst Appl Microbiol 38:472–482.

663 20. Ritari J, Salojärvi J, Lahti L, de Vos WM. 2015. Improved taxonomic assignment of human
664 intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 16:1056.

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

668 22. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. 2011. A Guide to the Natural
669 History of Freshwater Lake Bacteria. Microbiol and Mol Biol Rev 75:14-49.

670 23. Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD. 2018. TaxAss: Leveraging a Custom
671 Freshwater Database Achieves Fine-Scale Taxonomic Resolution. mSphere 3:1–14.

672 24. Newton IL, Roeselers G. 2012. The effect of training set on the classification of honey bee
673 gut microbiota using the Naïve Bayesian Classifier. BMC Microbiol 12:221.

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

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

680 27. Burke CM, Darling AE. 2016. A method for high precision sequencing of near full-length
681 16S rRNA genes on an Illumina MiSeq. PeerJ 4:e2492.

682 28. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK.
683 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-
684 nucleotide resolution. Nucleic Acids Res 47:e103.

685 29. Karst SM, Ziels RM, Kirkegaard RH, Albertsen M. 2019. Enabling high-accuracy long-read
686 amplicon sequences using unique molecular identifiers and Nanopore sequencing. bioRxiv
687 645903.

688 30. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. 2015. Back to basics - the
689 influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge
690 communities. PLoS One 10:e0132783.

691 31. Lane DJ. 1991. 16S/23S rRNA sequencing, p. 115–175. *In* Stackebrandt, E, Goodfellow, M
692 (eds.), Nucleic Acid Techniques in Bacterial Systematics. John Wiley and Sons, Chichester,
693 United Kingdom.

694 32. Massana R, Murray AE, Preston CM, DeLong EF. 1997. Vertical distribution and
695 phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel.
696 Appl Environ Microbiol 63:50–56.

697 33. Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. 2018. Newly designed 16S rRNA
698 metabarcoding primers amplify diverse and novel archaeal taxa from the environment.
699 Environ Microbiol Rep 11:487–494.

700 34. Tange O. 2018. GNU Parallel 2018. Ole Tange.

701 35. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
702 Bioinformatics 26:2460–2461.

703 36. Pruesse E, Peplies J, Glöckner FO. 2012. SINA: Accurate high-throughput multiple sequence
704 alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829.

705 37. Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2019. Biostrings: Efficient manipulation of
706 biological strings.

707 38. Corporation M, Weston S. 2019. doParallel: Foreach Parallel Adaptor for the "parallel"
708 Package.

709 39. Wickham H. 2019. stringr: Simple, Consistent Wrappers for Common String Operations.

710 40. Dowle M, Srinivasan A. 2019. data.table: Extension of `data.frame`.

711 41. Wickham H, Henry L. 2019. tidyr: Easily Tidy Data with "spread()" and "gather()"
712 Functions.

713 42. Wickham H, François R, Henry L, Müller K. 2019. dplyr: A Grammar of Data Manipulation.

714 43. Kirkegaard RH, McIlroy SJ, Kristensen JM, Nierychlo M, Karst SM, Dueholm MS,
715 Albertsen M, Nielsen PH. 2017. The impact of immigration on microbial community
716 composition in full-scale anaerobic digesters. Sci Rep 7:9343

717 44. Muyzer G, de Waal EC, Uitterlinden AG. 1993. Profiling of complex microbial populations
718 by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified

719 genes coding for 16S rRNA. AEM 59:695–700.

720 45. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees
721 for large alignments. PLoS One 5:e9490.

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

725 47. R Core Team. 2016. R: A language and environment for statistical computing. R Foundation
726 for Statistical Computing, Vienna, Austria.

727 48. RStudio Team. 2015. RStudio: Integrated Development Environment for R. Boston, MA.

728 49. Wickham H. 2009. ggplot2 - Elegant Graphics for Data AnalysisSpringer. Springer Science
729 & Business Media.

730 50. Andersen KS, Kirkegaard RH, Karst SM, Albertsen M. 2018. ampvis2: an R package to
731 analyse and visualise 16S rRNA amplicon data. bioRxiv 299537.

732 51. Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, Zhang Q, Brown M, Li Z, Van Nostrand
733 JD, Ling F, Xiao N, Zhang Y, Vierheilig J, Wells GF, Yang Y, Deng Y, Tu Q, Wang A,
734 Zhang T, He Z, Keller J, Nielsen PH, Alvarez PJJ, Criddle CS, Wagner M, Tiedje JM, He Q,
735 Curtis TP, Stahl DA, Alvarez-Cohen L, Rittmann BE, Wen X, Zhou J. 2019. Global diversity
736 and biogeography of bacterial communities in wastewater treatment plants. Nat Microbiol
737 4:1183-1195.

738 52. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz
739 P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially
740 revises the tree of life. Nat Biotechnol 36:996-1004.

741 53. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P,
742 Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially
743 expands the tree of life. Nat Microbiol 2:1533–1542.

744 54. Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from

745 microbiomes using nanopore sequencing. Nat Biotechnol. https://doi.org/10.1038/s41587-
746 020-0422-6

747 55. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH,
748 Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M. 2020. Connecting
749 structure to function with the recovery of over 1000 high-quality activated sludge
750 metagenome-assembled genomes encoding full-length rRNA genes using long-read
751 sequencing. bioRxiv 088096.

752 56. Isazadeh S, Jauffur S, Frigon D. 2016. Bacterial community assembly in activated sludge:
753 mapping beta diversity across environmental variables. MicrobiologyOpen 5:1050–1060.

754 57. Gonzalez-Martinez A, Rodriguez-Sanchez A, Lotti T, Garcia-Ruiz MJ, Osorio F, Gonzalez-
755 Lopez J, Van Loosdrecht MCM. 2016. Comparison of bacterial communities of conventional
756 and A-stage activated sludge systems. Sci Rep 6:18786.

757 58. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig
758 W, Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic
759 frameworks. Nucleic Acids Res 42:D643–D648.

760 59. McIlroy SJ, Karst SM, Nierychlo M, Dueholm MS, Albertsen M, Kirkegaard RH, Seviour
761 RJ, Nielsen PH. 2016. Genomic and in situ investigations of the novel uncultured Chloroflexi
762 associated with 0092 morphotype filamentous bulking in activated sludge. ISME J 10:2223-
763 2234.

764 60. Nierychlo M, Mcilroy SJ, Kucheryavskiy S, Jiang C, Ziegler AS, Kondrotaite Z, Stokholm-
765 Bjerregaard M, Nielsen PH. 2020. Candidatus Amarolinea and Candidatus Microthrix Are
766 Mainly Responsible for Filamentous Bulking in Municipal Danish Wastewater Treatment
767 Plants. Front Microbiol. doi: 10.3389/fmicb.2020.01214

768 61. Edgar RC. 2016. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv
769 074252.

770 62. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013.
771 Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-
772 generation sequencing-based diversity studies. Nucleic Acids Res 41:1–11.

773 63. Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA. 2019.
774 Species abundance information improves sequence taxonomy classification accuracy. Nat
775 Commun 10:1–10.

776 64. Peterson J, Bonazzi V, Starke-Reed P, Read J, Zakhari S, Baker CC, Garges S, Howcroft TK,
777 Wang L, Giovanni M, Grandison L, McEwen JE, Watson B, Lunsford RD, Akolkar B, Guyer
778 M, Giblin C, David H, Khalsa J, Humble M, Pontzer C, Mills M, Salomon R, Wright M,
779 Schloss JA, Begg L, Belachew T, Sayre MH, Karp RW, Davis C, Mullins C, Peavy H, Di
780 Francesco V, Wetterstrand KA, McInnes P, Wellington CR, Portnoy M, Little AR, Deal C.
781 2009. The NIH Human Microbiome Project. Genome Res 19:2317–2323.

782 65. Apprill A, Mcnally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA
783 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb
784 Ecol 75:129–137.

785 66. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: Assessing small subunit
786 rRNA primers for marine microbiomes with mock communities, time series and global field
787 samples. Environ Microbiol 18:1403–1414.

788 67. Bodenhausen N, Horton MW, Bergelson J. 2013. Bacterial Communities Associated with the
789 Leaves and the Roots of Arabidopsis thaliana. PLoS One 8:e56329.

790 68. Chelius MK, Triplett EW. 2001. The diversity of archaea and bacteria in association with the
791 roots of Zea mays L. Microb Ecol 41:252–263.

792 69. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
793 reads. EMBnet.journal 17:10–12.

794

795 **Tables:**

796 **Table 1: Numbers and percentage of FL-ASVs estimated to belong to novel taxa.** FL-ASVs
797 were mapped to SILVA 138 SSURef NR99 to find the identity with the closest relative in the
798 database. The novelty was determined based on the identity for each FL-ASV using the taxonomic
799 thresholds proposed by Yarza et al. (2014).

800

	Sequences	Percentage
Phylum (<75.0%)	1	0.01%
Class (<78.5%)	1	0.01%
Order (<82.0%)	3	0.03%
Family (<86.5%)	16	0.17%
Genus (<94.5%)	548	5.76%
Species (<98.7%)	2449	25.7%

801

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

	De novo taxa	Percentage
Phylum	1	2.22%
Class	12	10.17%
Order	66	24.35%
Family	276	47.18%
Genus	1324	72.91%
Species	3973	95.28%

803

Table S1: Oligonucleotides used for full-length 16S rRNA gene library preparation. Unique molecular tags, and sample barcodes are marked with blue and red, respectively. The GGGGGG motif marked with bold typefaces may form strong secondary structures known as guanine tetraplexes. This motif was replaced by GCGCGG in the DNA-based protocol, which increased the robustness of library preparation. Oligoes marked with "*" are identical between the RNA- and DNA-based protocols.

Name	Sequence	Protocol step
SSU_rRNA_RT2	A**GGGGGG**CAAAGATGAAGATNNNNNNNNNNNNNNNNCTAGTACGACTTGCCTGTCGCTCTATCTTCTTTTTTTTTTTTTTTTTTVN	First strand cDNA synthesis
SSU_rRNA_RT3	A**GGGGGG**CAAAGATGAAGATNNNNNNNNNNNNNNNNTTCTGCCTACTTGCCTGTCGCTCTATCTTCTTTTTTTTTTTTTTTTTTVN	
SSU_rRNA_s	5'Phos-GGGCAATATCAGCACCAACAGAAA-3'SpC3	Adaptor ligation
SSU_rRNA_l	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNNNNATAGAGGCTTTCTGTTGGTGCTGATATTGC	Second strand cDNA syntehsis
SSU_rRNA_pcr_fw*	CTCCACCCAGACTCATCCAT	Library and clonal amplification
SSU_rRNA_pcr_rv	A**GGGGGG**CAAAGATGAAGAT	
SSU_rRNA_readtag_fw*	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCACCCAGACTCATCCAT	Read-tag library preparation
SSU_rRNA_readtag_rv	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTA**GGGGGG**CAAAGATGAAGAT	
SSU_rRNA_linktag_fw	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAGATAGAGCGACAGGCAAGT	Linked-tag library preparation
SSU_rRNA_linktag_rv	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCAATATCAGCACCAACAGAAA	
SSU_rRNA_read2_fw*	GCTCTTCCGATCTCTCCACCCAGACTCATCCAT	Illumina MiSeq/HiSeq sequencing
SSU_rRNA_read2_rv	GCTCTTCCGATCTA**GGGGGG**CAAAGATGAAGAT	
f16S_rDNA_pcr1_fw1	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNNNNTGCCTCTTAGAGTTTGATCMTGGCTCAG	Adaptor tagging
f16S_rDNA_pcr1_fw2	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNNNNTCCTCTACAGAGTTTGATCMTGGCTCAG	
f16S_rDNA_pcr1_rw-v2	A**GCGCGG**CAAAGATGAAGATNNNNNNNNNNNNNNNNCGTACTAGTACGGYTACCTTGTTACGACTT	
f16S_rDNA_pcr2_fw*	CTCCACCCAGACTCATCCAT	Library and clonal amplification
f16S_rDNA_pcr2_rv-v2	A**GCGCGG**CAAAGATGAAGAT	
f16S_rDNA_readtag_fw*	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCACCCAGACTCATCCAT	Read-tag library preparation
f16S_rDNA_readtag_rv	CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTA**GCGCGG**CAAAGATGAAGAT	
f16S_rDNA_linktag_fw	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTGAGCCAKGATCAAACTCT	Linked-tag library preparation
f16S_rDNA_linktag_rv	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGTCGTAACAAGGTARCCGTA	
f16S_rDNA_read2_fw*	GCTCTTCCGATCTCTCCACCCAGACTCATCCAT	Illumina MiSeq/HiSeq sequencing
f16S_rDNA_read2_rv-v2	GCTCTTCCGATCTA**GCGCGGC**AAAGATGAAGAT	

811 **Table S2: Samples used in this study.**

Sample/SeqID	Environment	Location	Sampling_date	Latitude	Longitude	Design	PMID
MQ180605-111	WWTP	Fredericia	150824	55.552368	9.720404	EBPR	This study
MQ180605-112	WWTP	Randers	150831	56.453767	10.070249	EBPR	This study
MQ180605-113	WWTP	Avedoere	150902	55.607888	12.448949	EBPR	This study
MQ180605-114	WWTP	Damhusaaen	150831	55.641184	12.505800	EBPR	This study
MQ180605-115	WWTP	Viborg	150825	56.425251	9.455198	EBPR	This study
MQ180605-116	WWTP	Ribe	150824	55.329053	8.74336	EBPR	This study
MQ180605-117	WWTP	Skive	150825	56.565132	9.042158	EBPR	This study
MQ180605-118	WWTP	Ejby Moelle	150825	55.399534	10.414787	EBPR	This study
MQ180605-119	WWTP	Esbjerg W	150921	55.488097	8.430505	BNR	This study
MQ180605-120	WWTP	Aalborg E	150824	57.045161	10.045761	EBPR	This study
MQ180605-121	WWTP	Haderslev	150902	55.249786	9.508609	EBPR	This study
MQ180605-122	WWTP	Odense NE	150908	55.432604	10.458855	EBPR	This study
MQ180605-123	WWTP	Odense NW	150908	55.421534	10.366234	BNR	This study
MQ180605-124	WWTP	CP Kelco	150901	55.507272	12.145488	BNR	This study
MQ180605-125	WWTP	Aalborg W	150824	57.049513	9.864788	EBPR	This study
MQ180605-126	WWTP	Hjoerring	150831	57.421265	9.975411	EBPR	This study
MQ180605-127	WWTP	Egaa	150825	56.21314	10.242467	EBPR	This study
MQ180605-128	WWTP	Esbjerg E	150826	55.458597	8.489092	BNR	This study
MQ180605-129	WWTP	Bjergmarken	150825	55.649286	12.058568	EBPR	This study
MQ180605-130	WWTP	Boeslum	150921	56.198467	10.728742	EBPR	This study
MQ180605-131	WWTP	Mariagerfjord	150831	56.741769	10.115426	EBPR	This study
MQ180605-132	WWTP	Hirtshals	150914	57.577275	9.992971	EBPR	This study
16SAMP-4598	WWTP	Lundtofte	140623	55.802079	12.538763	EBPR	28496434
16SAMP-4601	WWTP	Ringkoebing	140625	56.089793	8.236342	EBPR	28496434
16SAMP-2340	WWTP	Aaby	100819	56.150696	10.176274	EBPR	28496434
16SAMP-2643	WWTP	Aars	100819	56.810414	9.547714	BNR	28496434
16SAMP-2589	WWTP	Fornaes	110819	56.437583	10.913985	BNR	28496434
MQ180605-133	Anaerobic digester	Bjergmarken	150826	55.6494	12.0589	Thermophilic	This study
MQ180605-134	Anaerobic digester	Ejby Moelle	150826	55.3981	10.415	Mesophilic	This study
MQ180605-135	Anaerobic digester	Fredericia	150826	55.5752	9.7476	THP + Mesophilic	This study
MQ180605-136	Anaerobic digester	Randers	150826	56.4697	10.0373	Mesophilic	This study
MQ180605-137	Anaerobic digester	Slagelse	150826	55.4163	11.3404	Mesophilic	This study
MQ180605-138	Anaerobic digester	Viborg	150826	56.4251	9.4543	Mesophilic	This study
MQ180605-139	Anaerobic digester	Aaby	150826	56.1512	10.1756	Thermophilic	This study
MQ180605-140	Anaerobic digester	Aalborg W	150826	57.0493	9.8635	Thermophilic	This study
MQ180605-141	Anaerobic digester	Esbjerg W	150826	55.4882	8.4307	Mesophilic	This study
MQ180605-142	Anaerobic digester	Damhusaaen	150902	55.6408	12.5059	Mesophilic	This study
MQ180605-143	Anaerobic digester	Fornaes	150902	56.4377	10.9145	Mesophilic	This study
MQ180605-144	Anaerobic digester	Hjoerring	150902	57.4196	9.9756	Mesophilic	This study
MQ180605-145	Anaerobic digester	Avedoere	150902	55.6087	12.4505	Mesophilic	This study
MQ180605-146	Anaerobic digester	Mariagerfjord	150902	56.741769	10.115189	Mesophilic	This study
MQ180605-147	Anaerobic digester	Naestved	150826	55.2111	11.7285	THP + Mesophilic	This study
MQ180605-148	Anaerobic digester	Soeholt	150826	56.1753	9.5826	Mesophilic	This study
16SAMP-3460	Anaerobic digester	Hobro	130828	56.6375	9.8085	Mesophilic	28839166
16SAMP-3463	Anaerobic digester	Lundtofte	130828	55.8024	12.5388	Mesophilic	28839166
16SAMP-4077	Anaerobic digester	Ringkoebing	130828	56.0882	8.2568	Mesophilic	28839166
16SAMP-4081	Anaerobic digester	Egaa	130501	56.21314	10.242467	Mesophilic	This study
16SAMP-3459	Anaerobic digester	Hobro	130821	56.1353	8.9731	NA	This study
16SAMP-4076	Anaerobic digester	Nysted	130821	54.699063	11.712259	NA	This study

* THP: thermal hydrolysis pretreatment; PMID: Pubmed ID reference

812

813 **Table S3: Classification of raw assemblies, full-length 16S rRNA, and FL-ASVs.**

Environment	Library type	Data set	nSeqs	A16S	A23S	B16S	B23S	E18S	E28S	Unclassified
Activated sludge	RNA based	Raw assemblies	274,240	2,730	306	119,947	129,851	6,139	13,689	1,540
Activated sludge	RNA based	Full-length 16S rRNA	45,907	396	0	45,496	0	15	0	0
Activated sludge	DNA based	Raw assemblies	506,519	0	0	506,518	0	0	0	1
Activated sludge	DNA based	Full-length 16S rRNA	468,426	0	0	468,425	0	0	0	1
Anaerobic digester	RNA based	Raw assemblies	361,091	73,583	8,990	79,057	196,916	0	206	2,339
Anaerobic digester	RNA based	Full-length 16S rRNA	38,834	16,494	0	22,340	0	0	0	0
Anaerobic digester	DNA based	Raw assemblies	402,232	0	0	402,232	0	0	0	0
Anaerobic digester	DNA based	Full-length 16S rRNA	373,340	0	0	373,340	0	0	0	0
Combined	Combined	FL-ASVs	9,521	75	0	9,446	0	0	0	0

814

Figures:

Figure 1:

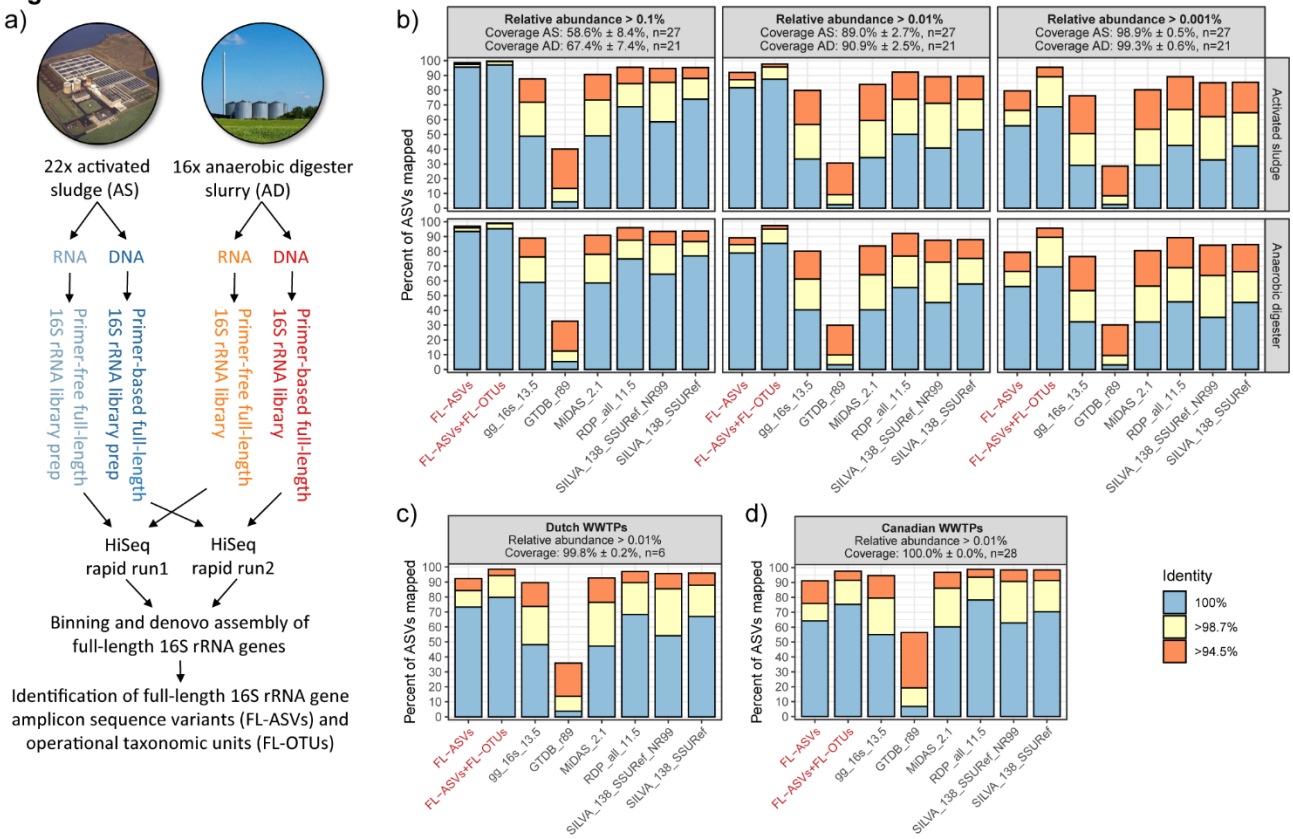

Figure 1. Construction and evaluation of the FL-ASV and FL-OTU expanded reference databases. a) Preparation of FL-ASVs and FL-OTUs. Samples were collected from WWTPs and ADs, and DNA and RNA were extracted. Purified DNA or RNA was used for the preparation of primer-based and "primer-free" full-length 16S rRNA libraries, respectively. These were sequenced and processed bioinformatically to produce the FL-ASVs and FL-OTUs b) Mapping of V1-3 amplicon data to the FL-ASV reference database, the FL-OTU expanded database, and commonly applied universal reference databases. ASVs were obtained from activated sludge and anaerobic digester samples and filtered based on their relative abundance before the analyses to uncover how well the databases cover the rare biosphere. The fraction of the microbial community represented by the remaining ASVs after the filtering (coverage) is shown as the mean ± standard deviation across plants. c) Mapping of V1-3 ASVs from Dutch WWTPs based on raw data from Gonzalez-Martinez et al. (57). For details, see Figure S2a. d) Mapping of V3-5 ASVs from Canadian WWTPs, based on raw data from Isazadeh et al. 2016 (56). For details, see Figure S2b.

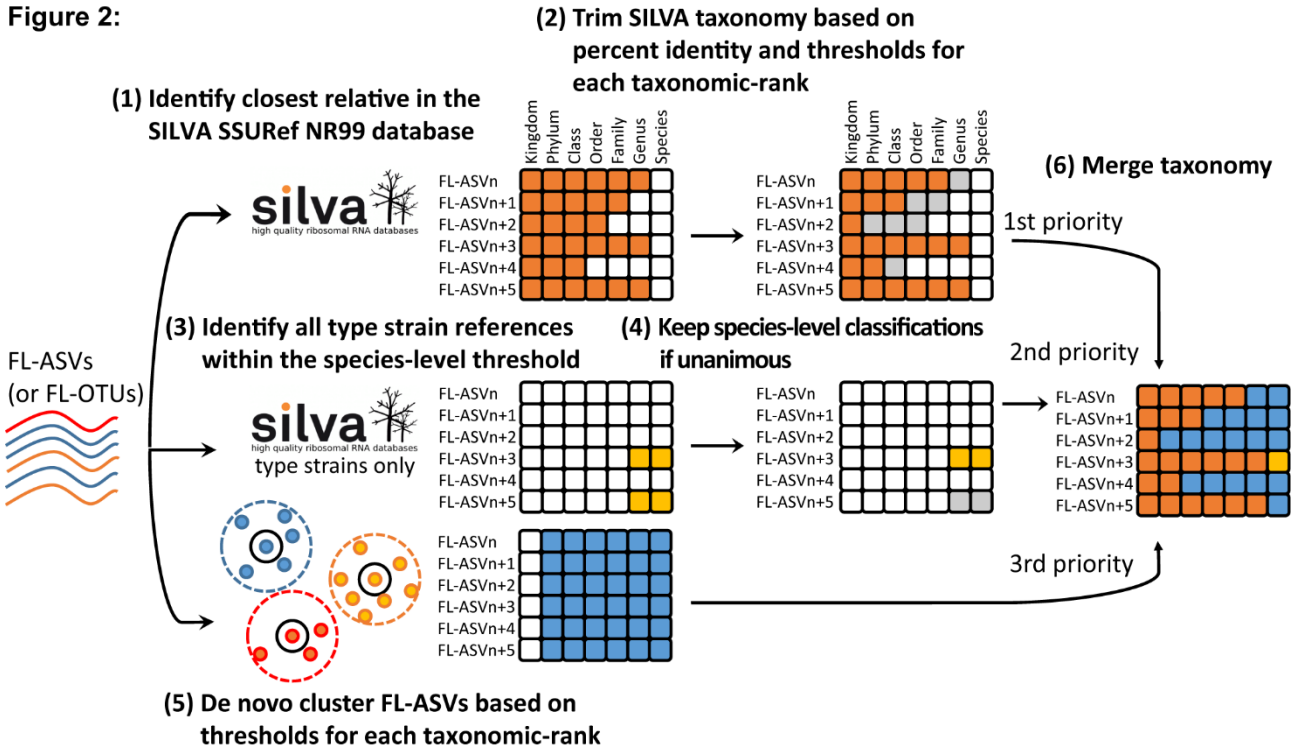

Figure 2. The AutoTax taxonomic framework. (1) FL-ASVs (or FL-OTUs) are first mapped to the SILVA 138 SSURef NR99 database to identify the closest relative and the shared percent identity. (2) Taxonomy is adopted from this sequence after trimming based on percent identity and the taxonomic thresholds proposed by Yarza *et al.* (12). To gain species-level information, FL-ASVs are also mapped to sequences from type strains extracted from the SILVA database, and species names are adopted if the identity is >98.7% and only a single species is within the threshold. (3) FL-ASVs are also clustered at different percent identity, corresponding to the thresholds proposed by Yarza *et al.* (12). The clustering is used to generate a stable *de novo* taxonomy. (4) Finally, a comprehensive taxonomy is obtained by filling gaps in the SILVA-based taxonomy with the *de novo*-taxonomy. Colored squares represent sources of taxonomic classifications of FL-ASVs: orange SILVA SSURef NR99, yellow: SILVA type strains, blue: *de novo* names, grey: names rejected during the AutoTax workflow.

Figure 3:

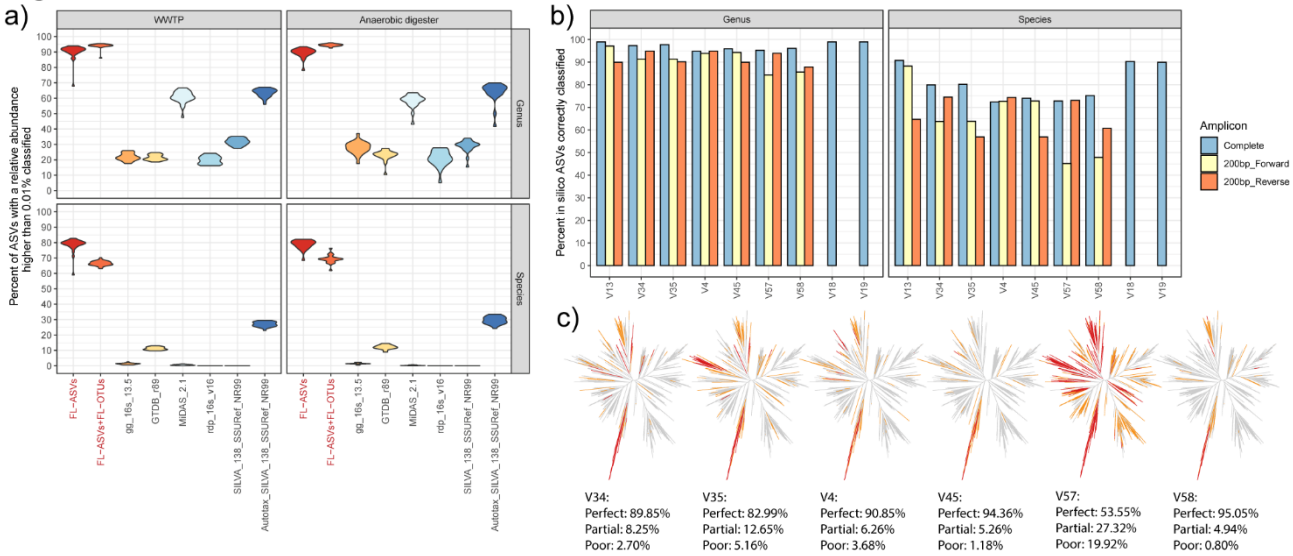

Figure 3. Evaluation of amplicon classification and primer specificity. a) The fraction of V1-3 ASVs from activated sludge and anaerobic digester sample with a relative abundance higher than 0.01% that were classified to the genus and species level using the FL-ASV reference database, the FL-OTU expanded database, and commonly applied universal reference databases, including the Autotax-processed SILVA 138 SSURef NR99 database. b) Classification of *in silico* bacterial ASVs, corresponding to amplicons produced using typical amplicon primers and the FL-ASV database. Results are shown for the complete amplicons as well as for partial amplicons corresponding to the first 200 bp from the forward or reverse read. c) Predicted primer bias associated with commonly used amplicon primers. The trees show the bacterial FL-ASVs, and the branch colors represent perfect hits (gray), partial hits (orange), and poor hits (red) (see materials and methods for definitions). The following primers were used: V13 (Lane 1991) (31), V18, V19, V34, and V58 (Klindworth et al. 2013) (62), V35 (Peterson et al. 2009) (64), V4 (Apprill et al. 2015) (65), V45 (Parada et al. 2016) (66), V57 (Chelius and Triplett 2001 and Bodenhausen et al. 2013) (67, 68).

Figure S1:

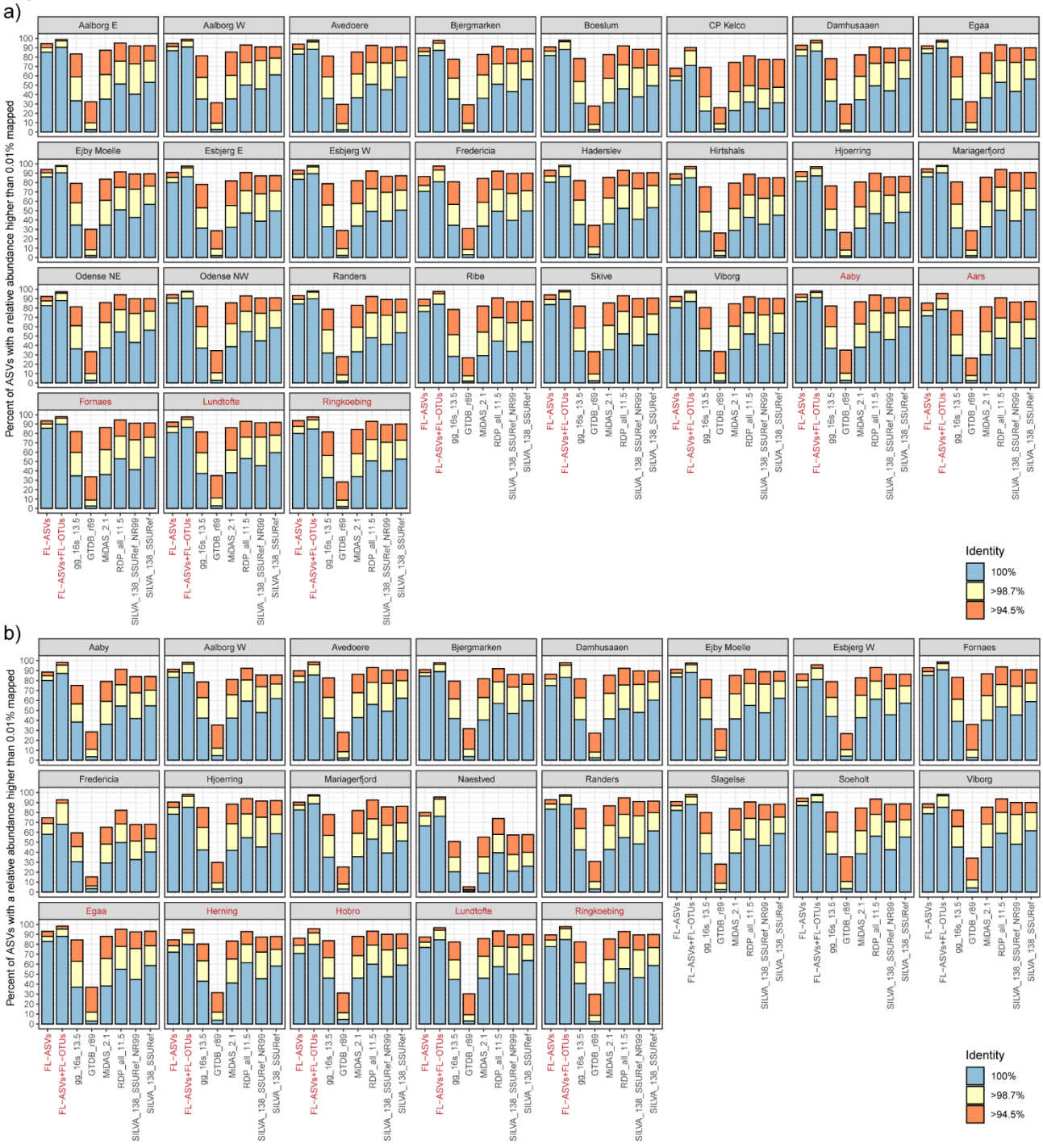

Figure S1: Mapping of activated sludge and anaerobic digester amplicon data to the FL-ASV and FL-OTU expanded reference database and public reference databases. V1-3 ASVs were obtained from the individual activated sludge (a) and anaerobic digester (b) samples used to create the FL-ASV reference database (black labels) and from other WWTPs (red labels).

Figure S2:

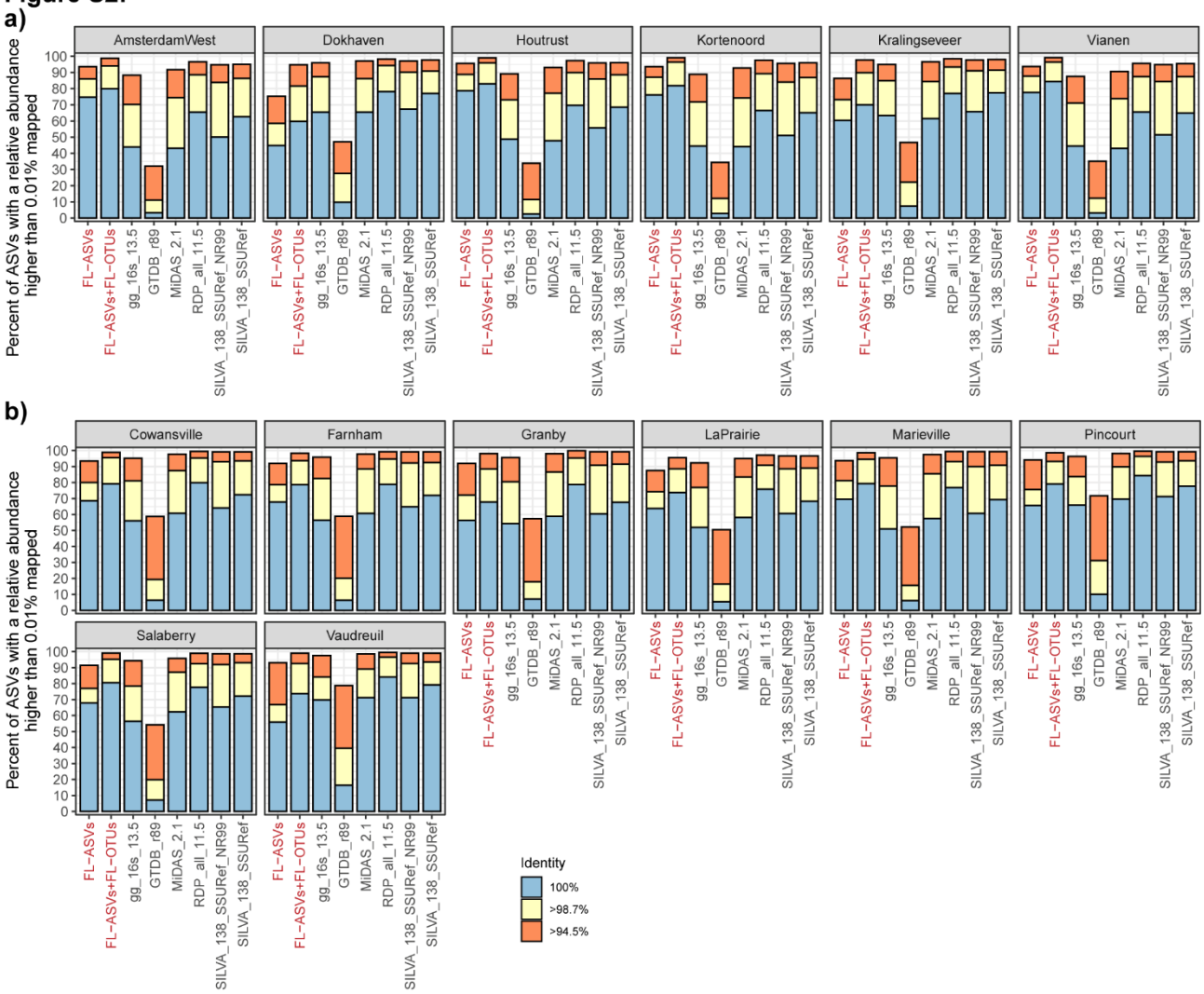

Figure S2: Mapping of amplicon data from non-Danish WWTP to the FL-ASV and FL-OTU expanded reference database and public reference databases. a) ASVs were created based on raw V1-3 amplicon reads obtained from Gonzarlez-Martinez *et al.* (57). Raw 454 genome sequencer reads in sff format were converted to fastq using CLC genomic workbench v. 20.0. Sequences were trimmed for the forward primer and truncated to 250 bp using *usearch -fastx_truncate -stripleft 19 -trunclen 250*, and quality filtered using *usearch -fastq_filter -fastq_maxee 1*. Processed reads were dereplicated using *usearch -fastx_uniques -sizeout*, denoised using *usearch -unoise3* (4). An ASV-table was produced with *usearch -otutab -zotus -sample_delim . -strand both*, and analyzed in R as described in the main article. All plants are based on conventional activated sludge except Dokhaven, which is configured with an Adsorption-Belebungsverfahren (AB) process. b) ASVs were created based on raw V3-5 amplicon reads obtained from Isazadeh *et al.* (2016)(56). Read 1 in SRA data were first oriented based on the

882 SILVA 138 SSURef NR99 database using *usearch -orient*. Sequences were trimmed for the
883 forward primer using *cutadapt -g CTACGGRNGGCWGC* (69), truncated to 250bp using *usearch -*
884 *fastx_truncate -trunclen 250*, and quality filtered using *usearch -fastq_filter -fastq_maxee 1*.
885 Processed reads were dereplicated using *usearch -fastx_uniques -sizeout*, denoised using *usearch -*
886 *unoise3* (4). An ASV-table was produced with *usearch -otutab -zotus -sample_delim . -strand both*,
887 and analyzed in R as described in the main article. The data originate from samples collected at
888 eight Canadian WWTPs, which among others, differs in influent composition and treatment
889 processes, for details see Isazadeh *et al.* (2016)(56). Marieville, Farnham, and Cowansville:
890 Oxidation ditch. LaPrairie, Granby, Pincount, Salaberry: Conventional aeration. Vaudreuil:
891 Sequencing batch reactor carrousel, which employs the principles of oxidation ditch.
892
893

Figure S3:

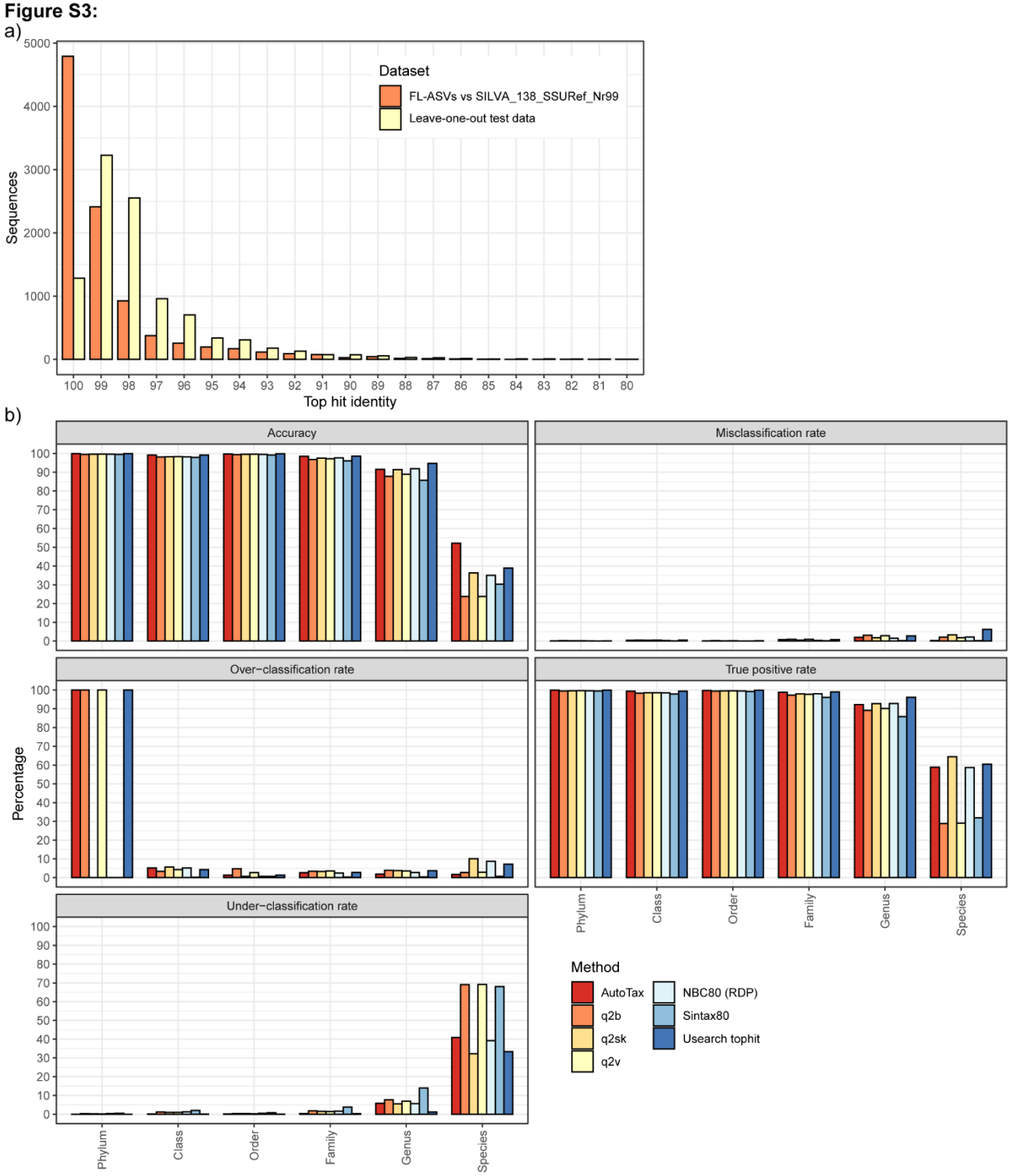

Figure S3. Evaluation of the taxonomy assignment to FL-ASVs. A leave-one-out validation test was performed using the AutoTax processed SILVA 138 SSURef NR99 database without the *de novo* taxonomy. This database only contains high-quality, full-length 16S rRNA sequences and improper-taxonomic entries (e.g., entries containing the words uncultured, unknown, incertae sedis, and possible) were removed. The database was divided into a 10,000 sequence test set, and the

900 remaining database which was used for classification. a) Top hit identity distribution of the leave-
901 one-out test data and for the FL-ASVs from this study mapped to the full SILVA 138 SSURef
902 NR99 database (real data). The distributions are in the same range, confirming that the test data
903 reflect actual data processed by AutoTax. b) Comparison of classification performance metrics for
904 the AutoTax taxonomy assignment and commonly applied classifiers. The metrics used were the
905 same as described by Edgar (13), but they were here applied to each taxonomic rank.
906 Accuracy=TP/(K+OC); Misclassification rate=MC/K; Over-classification rate=OC/L; True positive
907 rate=TP/K; Under-classification rate: UC/K, where K is the number of test sequences with a known
908 taxonomy, L the number of test sequences without a known taxonomy, TP the number of names
909 correctly assigned, MC the number of misclassification errors, OC the number overclassification
910 errors, and UC the number of underclassification errors. The following classifiers were used:
911 QIIME 2 feature-classifier with classify-consensus-blast (q2b), classify-consensus-vsearch (q2v),
912 and classify-sklearn (q2sk) (5), USEARCH v.11 with SINTAX, NBC (Reimplementation of RDP),
913 and USEARCH top hit (default parameters) (7, 13).
914

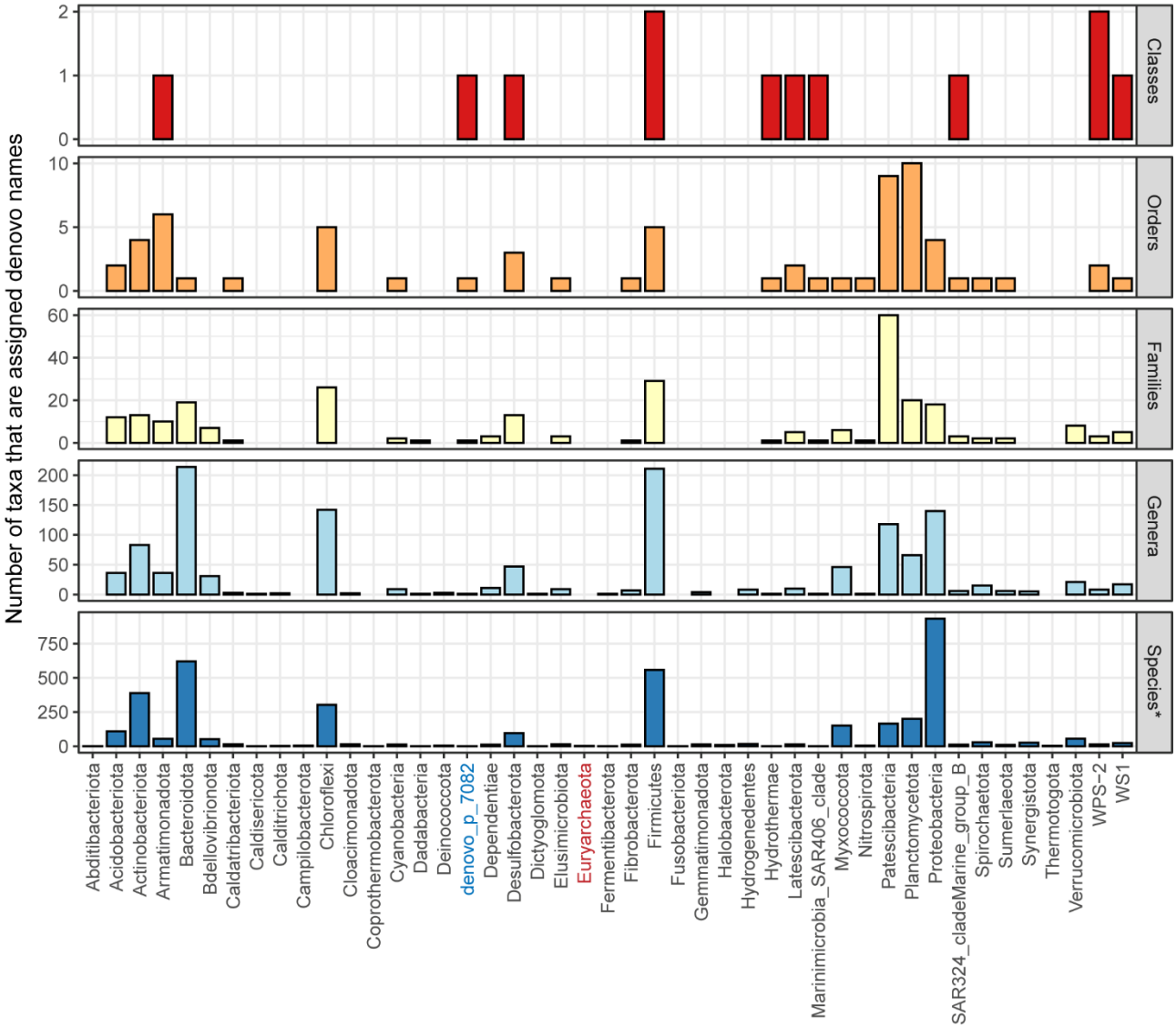

Figure S4. Novel taxa classified in the FL-ASV reference database. The number of taxa, which obtained names from the *de novo* taxonomy in different phyla. Archaeal phyla are highlighted in red and *de novo* phyla in blue.

Figure S5:

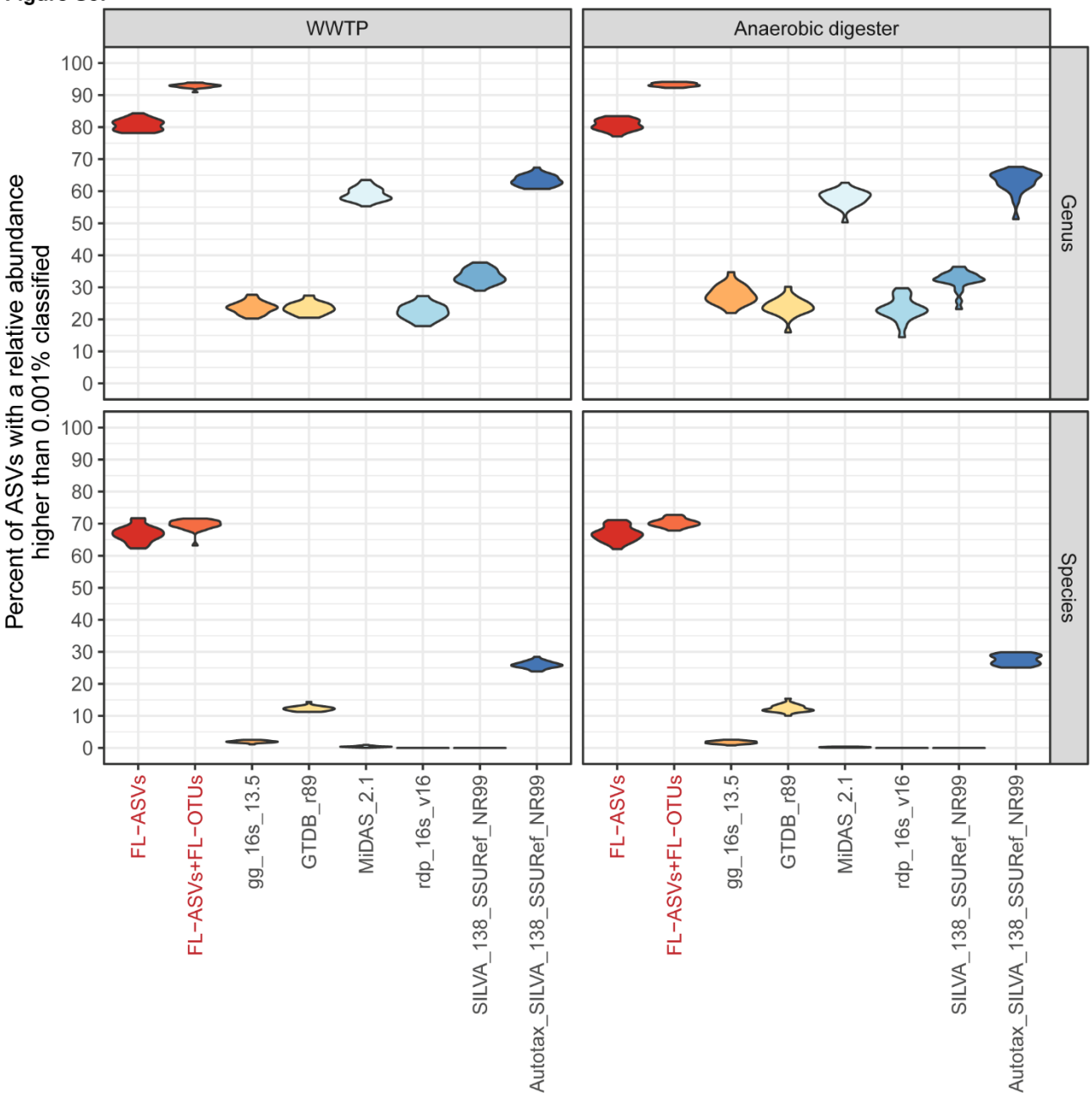

Figure S5. Effect of the rare biosphere on amplicon classification. a) The fraction of V1-3 ASVs from activated sludge and anaerobic digester sample with a relative abundance higher than 0.001% which was classified at the genus and species level using the FL-ASV reference database, the FL-OTU expanded database, and commonly applied universal reference databases, including the Autotax-processed SILVA 138 SSURef NR99 database.

Figure S6:

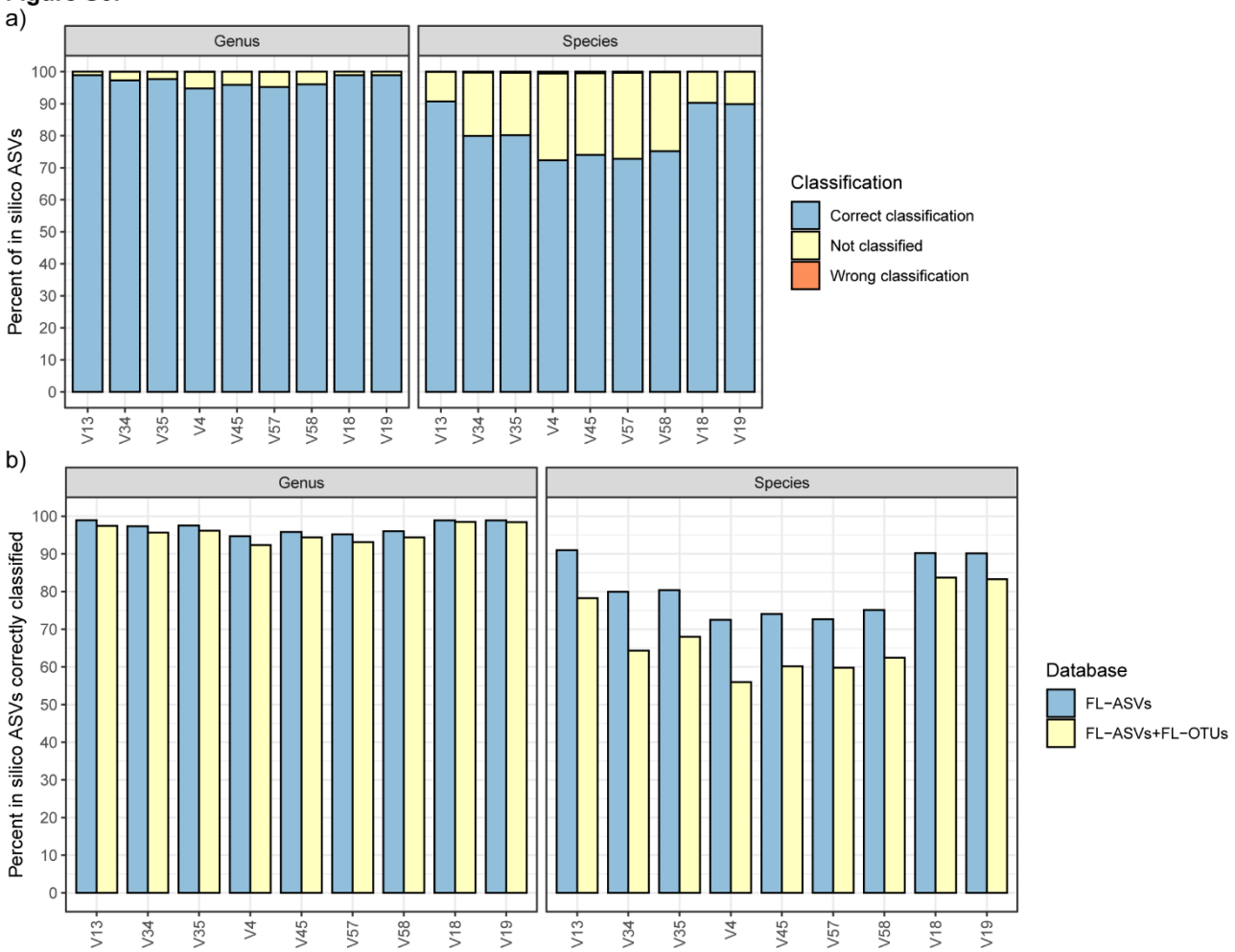

Figure S6. The specificity of taxonomy assignment using SINTAX and the FL-ASV and FL-OTU expanded reference database. a) Percentage of *in silico* bacterial ASVs, corresponding to amplicons produced using typical amplicon primers, that was correctly, wrongly, or not classified based on the FL-ASV database. b) Determining the effect of chimeras in the reference database. *In silico* bacterial ASVs were classified using the FL-ASV database and the FL-OTU expanded reference database.

936 **SUPPLEMENTARY INFORMATION FOR**

937

938 **Generation of comprehensive ecosystems-specific reference databases with**
939 **species-level resolution by high-throughput full-length 16S rRNA gene**
940 **sequencing and automated taxonomy assignment (AutoTax)**

941 Morten Simonsen Dueholm, Kasper Skytte Andersen, Simon Jon McIlroy, Jannie Munk Kristensen,
942 Erika Yashiro, Søren Michael Karst, Mads Albertsen, and Per Halkjær Nielsen

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

945

946

947 **Detailed description of the AutoTax procedure:**

948 *Full-length 16S rRNA exact amplicon sequence variant (FL-ASV) calling*

949 Full-length 16S rRNA gene sequences are first oriented according to the SILVA 138 SSURef Nr99
950 database (1) using the *usearch -orient* command. The sequences are then dereplicated using *usearch*
951 *-fastx_uniques* with the *-sizeout, -strand plus*, and *-threads 1* options. The *-sizeout* option put a size
952 annotation (how many times a unique sequence was observed) in the FASTA header. The *-strand*
953 *plus* option ensures that only correctly oriented sequences are considered when identifying identical
954 sequences. The dereplicated sequences are sorted and numbered based on the size annotation. The -
955 *threads 1* option ensures that sequences with the same size are always sorted and numbered the
956 same way in the output. The dereplicated sequences are finally denoised to produce the FL-ASVs
957 using the *usearch -unoise3* command with the *-minsize 2* option. The *-minsize* options specify the
958 minimum abundance required for detecting an ASV. The default value for short-read amplicons is
959 8; however as the synthetic long-read sequences are independently amplified (due to the unique
960 molecular identifiers (UMIs) added before the PCR steps) and have a very low error-rate we can
961 lower the threshold to 2 with little risk of introducing sequences with errors, see "Calculation of FL-
962 ASV error-rate" below.

963

964 *Preparation of chimera-filtered full-length 16S rRNA OTUs (FL-OTUs)*

965 Dereplicated sequences from above are clustered at 99% sequence identity to create FL-OTUs using
966 *usearch -cluster_smallmem* command with the *-id 0.99, -maxrejects 0, -centroids,* and *-sortedby*
967 *size* options. The *-maxrejects 0* option specified a complete database search, thereby providing more
968 confident clusters. Potential chimeras are identified and extracted using the *usearch -uchime2_ref*
969 command with the *-strand plus, -mode sensitive*, and -chimeras options using the FL-ASVs as a
970 reference database. The sensitive mode was chosen as it captures more chimeras (on the expense of
971 more false positives), as these are detrimental to the final reference database. The FL-ASVs were
972 used as the reference database because it contains exact references for the abundant 16S rRNA
973 genes, which are most likely to form chimeras. Chimeras are finally removed to create the chimera-
974 filtered FL-OTUs using the *usearch -search_exact* command with the *-strand plus -dbnotmatched*
975 options, using the identified chimera sequences as the query, and the pre-filtered FL-OTUs as the
976 reference database.
977
978 *Taxonomy assignment*
979 For the taxonomy assignment, we first create two independent taxonomies. The first taxonomy is
980 based on the most recent version of the SILVA SSURef Nr99 database and reflects the current state
981 of microbial taxonomy. The second is a robust, although not necessarily evolutionary correct, *de*
982 *novo* taxonomy. The latter will be used as a taxonomic placeholder for taxonomic ranks without
983 information in the SILVA-based taxonomy.
984
985 The first step in both cases is the alignment of each FL-ASV to the global alignment in the SILVA
986 reference database in ARB-format using SINA. The aligned sequences are hereafter trimmed to
987 position 1048 to 41788 in the global SILVA alignment using the generic Linux command awk. The
988 purpose of this trimming is two-fold. Firstly, it ensures that flanking sequences do not result in
989 artificially low identities when the FL-ASVs are mapped to reference sequences in the SILVA
990 database produced with the popular 27F and 1391R primers (2). Secondly, it allows a more robust
991 *de novo* clustering (3). After trimming, gaps are removed from the alignment of FL-ASVs using the
992 *usearch -fasta_stripgaps* command. Finally, the FL-ASVs are sorted based on the FL-ASV
993 numbers in R.
994
995 The SILVA based taxonomy is obtained by mapping each of the trimmed FL-ASVs to the most
996 recent version of the SILVA SSURef Nr99 and the type strain database in FASTA format and

997 obtaining the taxonomic information from the closest relative as well as the percent identity.
998 Mappings are done using the *usearch -usearch_global* command with the *-maxrejects 0, -*
999 *maxaccepts 0, -top_hit_only, -strand plus, -id 0,* and *-blast6out* options for the complete SILVA
1000 database and the *-maxrejects 0, -maxaccepts 0, -strand plus, -id 0.987,* and *-blast6out* for the type
1001 strains. The *-maxrejects 0* and *-maxaccepts 0* options ensures that a comprehensive search is
1002 performed. The *-top_hits_only* and *-id 0* options provide the best hit in the complete SILVA
1003 database, whereas the *-id 0.987* without the *-top_hit_only* option provides all hits within the species-
1004 level threshold. The output files from the SILVA reference mapping are loaded into R, and a data
1005 frame with columns for FL-ASV number, percent identity, and SILVA taxonomy of the closest
1006 relative is created. The taxonomy field is then split into the seven main taxonomic ranks from
1007 kingdom to species. Fields that contain any of the following whole words are cleared: "uncultured",
1008 "unknown", "unidentified", "incertae sedis", "metagenome", "bacterium", and "possible". In
1009 addition, we replace all "candidatus" with "Ca" and all white spaces with underscores. Finally, all
1010 characters except letters, numbers and period, dash, and underscore are removed.
1011
1012 The taxonomy obtained from the closest relative in the SILVA database does not necessarily match
1013 that of the FL-ASV. The taxonomy is therefore trimmed based on the percent identity between the
1014 FL-ASV and its closest relative. For this trimming we use the taxonomic thresholds proposed by
1015 Yarza *et al.* (4) (94.5% for genus-level, 86.5% for family-level, 82.0%, for order-level, 78.5% for
1016 class-level, and 75.0% for phylum-level). Species-level classifications are only obtained from the
1017 mapping against type strains, and no classifications are provided if reference sequences from more
1018 than one species are within the species-level threshold.
1019
1020 To generate a comprehensive *de novo* taxonomy, FL-ASVs are clustered based on the taxonomic
1021 thresholds proposed by Yarza *et al.* (4) using the *usearch -cluster_smallmem* command with the *-id*
1022 *x, -maxrejects 0, -uc,* and *-sortedby other* options. Where x represent the threshold for the given
1023 taxonomic rank. The cluster_smallmem algorithm with the *-sortedby other* option clusters the FL-
1024 ASVs based on when they appear in the input FASTA file, and the same clusters are therefore
1025 formed even though additional FL-ASVs are appended to the FL-ASV-database in the future. We
1026 confirmed this by processing only the first half of the FL-ASVs from this study, which provided
1027 identical clustering. The *-maxrejects 0* ensures that the best reference is found among all clustered
1028 sequences. The output file is a UCLUST-format tabbed text, which describes the clusters formed.

The six UCLUST output files (species- to phylum-level) are loaded into R, and each are converted into a data frame with two columns. The first column with clusters is named based on the taxonomic clustering rank, and the second column with the input sequences is named as the taxonomic rank below. Subsequently, the data frames are merged from species to phylum level. This results in a comprehensive taxonomy, where the clustered centroids determine affiliation to the taxonomic ranks above.

The SILVA-based taxonomy and the *de novo* taxonomy are finally merged in R by replacing empty fields in the SILVA-based taxonomy with the *de novo* taxonomic information. This results in a comprehensive taxonomy, where all FL-ASVs are described on all seven taxonomic ranks.

Merging of the two taxonomies results in a few cases where a taxon has more than one parent (e.g., sequences from the same species affiliate more than one genus). In these cases, the taxonomy of the FL-ASV with the lowest ASV-number within the taxa is assigned to all members. A log file is written with details of all such cases.

Calculation of FL-ASV error-rate:

The Unoise3 algorithm predicts which sequences represent true biological variants (ASVs) based on their relative abundance compared to those of closely related sequences with one or a few errors. Based on the error-rate profile of full-length 16S rRNA genes found in Karst et al. (5) and the assumption that sequencing-errors are randomly distributed we can calculate the relative chance of obtaining a correct ASV compared to an ASV with a single sequencing error when we only consider sequence seen at least twice:

The probability of obtaining a correct sequence for FL-ASV$_n$ is the product of the probability of obtaining a sequence without any errors and relative abundance of FL-ASV$_n$, and for the sequence to be seen twice, we need to square this number.

$$P_{correct}(FL.ASV_n) = \left(P_{no\ error} \times P_{FL.ESV_n}\right)^2$$

The probability of obtaining a sequence with a single error for FL-ASV$_n$ is the product of the probability of obtaining a sequence with one error and relative abundance of FL-ASV$_n$. However, to calculate the probability of picking a sequence with the same error, we need to divide this probability with three times the length of FL-ASV$_n$ because the error could occur at any position in the sequence and could be any of the three wrong nucleotides. The probability for seeing a specific incorrect sequence twice therefore becomes:

$$P_{one\ error}(FL.ASV_n) = P_{one\ error} \times P_{FL.ASV_n} \times \frac{P_{one\ error} \times P_{FL.ASV_n}}{3 \times length_{FL.ASV_n}(bp)}$$

The relative chance of obtaining a correct FL-ASV compared to an FL-ASV with a single sequencing error is therefore:

$$\frac{P_{correct}(FL.ASV_n)}{P_{one\ error}(FL.ASV_n)} = \frac{{P_{no\ error}}^2 \times 3 \times length_{FL.ASV_n}(bp)}{{P_{one\ error}}^2} = \frac{0.112^2 \times 3 \times 1450}{0.209^2} = 1249$$

The results show that for each time we obtain an FL-ASV with a single error, we should expect approximately 1250 times more sequences of the correct FL-ASV. The difference should be large enough for Unoise3 to be able to remove almost all sequences with errors even though the errors are not entirely randomly distributed.

1080 To confirm this, we reanalyzed the mock data used in Karst *et al.* (5). This data contained 7816
1081 trimmed sequences (27f to 1492r) from a balanced mixture of DNA from *Escherichia coli* MG1655
1082 (NC_000913), *Bacillus subtilis* subsp. subtilis str. 168 (NC_000964) and *P. aeruginosa* PAO1
1083 (NC_002516). Denoising with *usearch -unoise3* command and the *-minsize 2* option produced 16
1084 error-corrected sequences, of which 13 mapped perfectly to the 16S rRNA reference sequences, two
1085 mapped with a single error, and one mapped with five errors. As the two FL-ASVs with single
1086 errors were observed 14 and 4 times, respectively, it is highly unlikely that they represent
1087 sequencing error based on the previous error-rate calculations. The same is holds for the FL-ASV
1088 with five errors, although this sequence was observed only twice.
1089
1090 **References:**

1091 1. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013.
1092 The SILVA ribosomal RNA gene database project: Improved data processing and web-based
1093 tools. Nucleic Acids Res 41:D590-6.
1094 2. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013.
1095 Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-
1096 generation sequencing-based diversity studies. Nucleic Acids Res 41:1–11.
1097 3. Schloss PD. 2013. Secondary structure improves OTU assignments of 16S rRNA gene
1098 sequences. ISME J 7:457–460.
1099 4. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB,
1100 Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and
1101 uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol
1102 12:635–645.
1103 5. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. 2018.
1104 Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences
1105 without primer bias. Nat Biotechnol 36:190–195.

1106
1107