AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Applied and Environmental Science

Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax)

Morten Simonsen Dueholm,[a] Kasper Skytte Andersen,[a] Simon Jon McIlroy,[a]* Jannie Munk Kristensen,[a] Erika Yashiro,[a] Søren Michael Karst,[a] Mads Albertsen,[a] Per Halkjær Nielsen[a]

[a]Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

Morten Simonsen Dueholm and Kasper Skytte Andersen contributed equally to the work. Author order was determined based on who wrote the paper.

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) resolved 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. Reference databases processed with AutoTax greatly improves the classification of short-read 16S rRNA ASVs at the genus- and species-level, compared with 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.

KEYWORDS 16S RNA, gene sequencing, microbial communities, microbial ecology, taxonomy, wastewater treatment

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).

The 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 to 500 bp long) is amplified by PCR and sequenced. The amplicons are then clustered into operational taxonomic units (OTUs)

Citation Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, Albertsen M, Nielsen PH. 2020. Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax). mBio 11:e01557-20. https://doi.org/10.1128/mBio.01557-20.

Editor Nicole Dubilier, Max Planck Institute for Marine Microbiology

Address correspondence to Morten Simonsen Dueholm, md@bio.aau.dk, or Per Halkjær Nielsen, phn@bio.aau.dk.

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

Received 13 June 2020
Accepted 18 August 2020
Published 22 September 2020

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 the microbial community structure to predict ecosystem functions or process performance in engineered systems. First, without taxonomic assignment, it is not possible to compare results across studies that have used primers targeting different regions of the 16S rRNA gene. Second, short-read ASVs alone do not contain enough evolutionary information to resolve their phylogeny confidently (11, 12). This limitation 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 (high-quality minimum information about a metagenome-assembled genome [MIMAG] 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.

Taxonomic assignment to ASVs relies on a classifier, e.g., the SINTAX (7) or the RDP classifier (6), which uses statistical algorithms to compare each ASV to a full-length 16S rRNA gene reference database to propose the best estimate of their taxonomy. Confident classification at the lowest taxonomic ranks (genus and species) requires high-identity reference sequences ($\geq$98.7% identity) and a complete seven-rank taxonomy for all references (13). Neither of these criteria is met with the most commonly applied universal reference databases, e.g., Greengenes (14), SILVA (15), and RDP (16), as they lack sequences or taxonomic assignment for many uncultivated environmental taxa. Given the vast diversity predicted for microbial life on Earth (17), it will be some time before reference sequences for all species are generated, and the manual curation of their taxonomy will not be feasible.

A potential solution to the problems mentioned above is to create ecosystem-specific reference databases. They can either be ecosystem-curated versions of universal databases or smaller independent databases that only include sequences from the specific ecosystem. The MiDAS 2.0 database for microbes in biological wastewater treatment systems (18) and the Dictyopteran gut microbiota reference Database (DictDb) (19) are examples of ecosystem-curated versions of the SILVA databases, where the taxonomies for the abundant and process-critical microorganisms are manually curated and maintained. However, the mostly manual taxonomic curation of central reference databases is time-consuming and subjective. The universal reference databases are also clustered at 99% identity or below to keep the unique sequences to a manageable number, which reduces the taxonomic resolution.

Examples of independent ecosystem-specific databases include the human intestinal tract 16S taxonomic database (HITdb) (20), the human oral microbiome database (HOMD) (21), the freshwater-specific FreshTrain database (22, 23), the honey bee gut microbiota database (24), and the rumen and intestinal methanogen database (25). While such databases have been shown to improve the rate of classifications for amplicons, they generally contain a relatively limited number of sequences and are therefore associated with an inherent risk of over- or misclassification if the sequence being classified is not represented in the database. To address this issue, Rohwer et al. introduced the TaxAss algorithm that classifies amplicons using two reference databases, namely, a universal database and a small ecosystem-specific database (23).

Amplicons are first mapped to the ecosystem-specific database to determine the percent identity with the best hit, and those above a user-defined threshold are classified using the ecosystem-specific database. The remaining sequences are classified using the more extensive universal database. While higher rates of classification were achieved, an issue with the approach is the potential for closely related sequences that fall on either side of the user-defined threshold to receive very different taxonomies, especially if the ecosystem-specific database is not updated to reflect the evolving taxonomy of the universal reference database. As such, while current strategies to create ecosystem-specific databases have shown promise, there are critical issues that need to be resolved before their potential can be realized.

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

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

RESULTS AND DISCUSSION

Sampling and high-throughput sequencing of full-length 16S rRNA sequences. To obtain 16S rRNA gene reference sequences for Danish wastewater treatment plants (WWTPs) and anaerobic digesters (ADs), we sampled biomass from 22 typical WWTPs and 16 ADs treating waste activated sludge located at Danish wastewater treatment facilities (see Table S2 in the supplemental material). These facilities represent an important engineered ecosystem containing complex microbial communities of both bacteria and archaea, with the vast majority of microbes being uncultured and poorly characterized (30).

DNA and RNA were extracted and used for synthetic long-read 16S rRNA gene sequencing using both a primer-based and primer-free approach (26) (Fig. 1a). A total of 926,507 full-length 16S rRNA gene sequences were obtained after quality filtering (see Table S3 in the supplemental material). They were denoised with UNOISE3 to generate a comprehensive reference database of 9,521 FL-ASVs with an error rate below the detection limit according to theoretical calculations and an analysis of 7,816

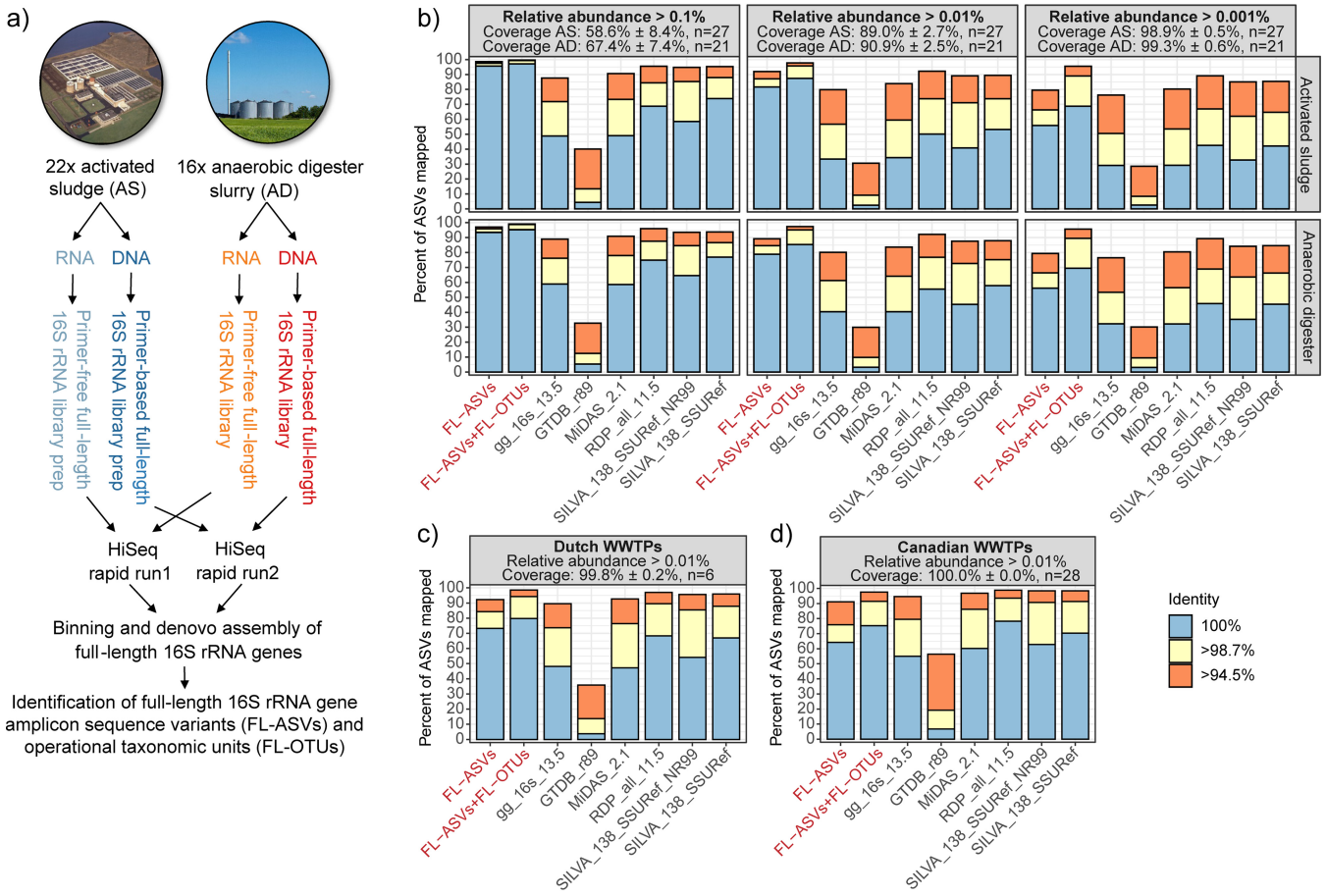

FIG 1 Construction and evaluation of FL-ASV and FL-OTU expanded reference databases. (a) Preparation of FL-ASVs and FL-OTUs. Samples were collected from WWTPs and ADs, and DNA and RNA were extracted. Purified DNA or RNA was used for the preparation of primer-based and "primer-free" full-length 16S rRNA libraries, respectively. They were sequenced and processed bioinformatically to produce the FL-ASVs and FL-OTUs. (b) Mapping of V1-V3 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-V3 ASVs from Dutch WWTPs based on raw data from Gonzalez-Martinez et al. (36). For details, see Fig. S2a. (d) Mapping of V3-V5 ASVs from Canadian WWTPs, based on raw data from Isazadeh et al. (35). For details, see Fig. S2b.

sequences previously obtained from the eight-strain ZymoBiomics microbial community DNA standard (26) (see Text S1 in the supplemental material).

To estimate the number of FL-ASVs belonging to novel taxa, FL-ASVs were mapped to the SILVA 138 SSURef NR99 database (15) using global mapping with USEARCH, and the identity of their closest relatives was compared with the thresholds for taxonomic ranks proposed by Yarza et al. (12) (Table 1). The majority of the FL-ASVs (94%) had references in the SILVA database above the genus-level threshold (>94.5% identity),

TABLE 1 Numbers and percentages of FL-ASVs[a] estimated to belong to novel taxa

Taxonomic rank (threshold)	No. of sequences	Percentage
New phylum (<75.0% identity)	1	0.01
New class (<78.5% identity)	1	0.01
New order (<82.0% identity)	3	0.03
New family (<86.5% identity)	16	0.17
New genus (<94.5% identity)	548	5.76
New species (<98.7% identity)	2,449	25.7

[a]FL-ASVs were mapped to SILVA 138 SSURef NR99 to find the identity with the closest relative in the database. The novelty was determined based on the identity for each FL-ASV using the threshold for each taxonomic rank proposed by Yarza et al. (12).

but 26% lacked references above the species-level threshold (98.7% identity), which are crucial for confident taxonomy assignment to ASVs.

FL-ASVs have better ecosystem coverage than universal reference databases. To evaluate if the FL-ASV database contained high-identity references for all bacteria in the ecosystem, we mapped V1-V3 ASVs obtained from the following two sources: (i) the same samples used to create the FL-ASVs and (ii) samples from unrelated Danish WWTP and ADs. The ecosystem-specific FL-ASV database (9,521 seq.) included more high-identity references (>98.7% identity) for the abundant ASVs (relative abundance cutoff at 0.01%) in all samples analyzed than that of the much larger universal databases MiDAS 2.1 (548,447 seq.) (18), SILVA 138 SSURef NR99 (510,984 seq.) (15), SILVA 138 SSURef (2,225,272 seq.) (15), Greengenes 16S v.13.5 (1,262,986 seq.) (14), and the full RDP v.11.5 (3,356,808 seq.) (16) (Fig. 1b and Fig. S1 in the supplemental material). ASVs were also mapped to the 16S rRNA gene database derived from the Genome Taxonomy Database (GTDB) release 89 (17,460 seq.) (31). However, this database lacked high-identity references for almost all ASVs. The poor coverage likely relates to the fact that 16S rRNA genes often fail to assemble in MAGs produced by short-read sequencing data (32). This problem will likely disappear in the future with the introduction of more high-quality MAGs with complete rRNA genes into the GTDB as a result of long-read sequencing technologies, such as Nanopore and PacBio (33, 34).

When the rare biosphere was included in the analysis (relative abundance cutoff at 0.001%), a decrease in the percentage of ASVs with high-identity reference sequences was observed (Fig. 1b). This may be a problem in ecosystems with high diversity, such as soil and sediments (9), where low-abundant microbes constitute a considerable fraction of the community, but also in engineered systems where transient or low-abundant bacteria, such as pathogens or bacteria degrading micropollutants, may be important. To get a better representation of the rare biosphere, we created an additional reference database, which besides the FL-ASVs, contained chimera-filtered, full-length 16S rRNA gene sequences clustered at 99% identity (FL-OTUs). This database greatly increased the coverage for the rare biosphere because it includes FL-OTUs for sequences that were only observed once. However, the improved coverage is achieved at the expense of taxonomic resolution (see later).

Since only Danish WWTPs and ADs were used to establish the comprehensive high-identity FL-ASV reference database, published amplicon data from non-Danish WWTPs (35, 36) were also evaluated (Fig. 1c and d and Fig. S2 in the supplemental material). Compared with the analyzed universal reference databases, the Danish reference FL-ASVs performed better or as well for most of the investigated non-Danish WWTPs, which indicates that even with less than 10,000 sequences, the database covers many of the microbes that are common to WWTPs across the world, particularly for systems with nutrient removal. We anticipate that our ongoing sampling of more than 1,000 WWTPs and AD systems across all 7 continents and different process-configurations (MiDAS Global, https://www.midasfieldguide.org/global) for high-throughput full-length 16S rRNA gene sequencing will provide references for the region-specific taxa in the near future, providing a comprehensive database of reference sequences for this ecosystem.

A new comprehensive taxonomic framework. A major limitation in the classification of amplicon data from environmental samples is the lack of genus and species names for many uncultivated bacteria in the universal reference databases. To address this limitation, we developed a simple taxonomic framework (AutoTax), which provides a consistent taxonomic classification of all reference sequences to all seven taxonomic ranks using identity thresholds (Fig. 2).

In the AutoTax pipeline, the FL-ASVs (or FL-OTUs) are first mapped to the SILVA 138 SSURef NR99 database, which provides the taxonomy and percent identity of the closest relative in the database. This taxonomy is assigned to the FL-ASV down to the taxonomic rank supported by the sequence identity thresholds proposed by Yarza et al. (12) (Table 1). Because of the stringent mapping and the taxonomy trimming based on percent identity, we obtained an overall better taxonomy assignment than that of

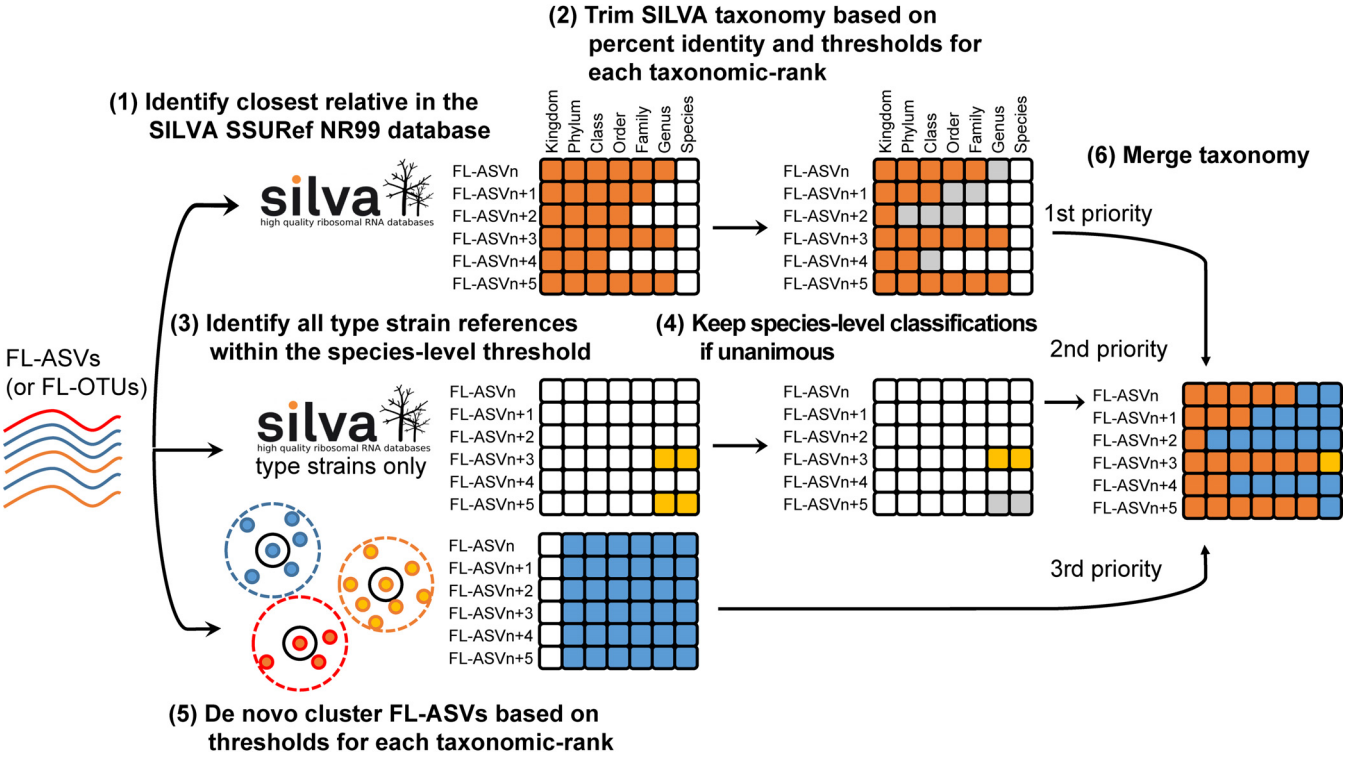

FIG 2 The AutoTax taxonomic framework. (1) FL-ASVs (or FL-OTUs) are first mapped to the SILVA 138 SSURef NR99 database to identify the closest relative and the shared percent identity. (2) Taxonomy is adopted from this sequence after trimming based on percent identity and the taxonomic thresholds proposed by Yarza et al. (12). (3) To gain species-level information, FL-ASVs are also mapped to sequences from type strains extracted from the SILVA database, and (4) species names are adopted if the identity is >98.7% and only a single species is within the threshold. (5) FL-ASVs are also clustered at different percent identities, corresponding to the thresholds proposed by Yarza et al. (12). The clustering is used to generate a stable *de novo* taxonomy. (6) 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, as follows: orange, SILVA SSURef NR99; yellow, SILVA type strains; blue, *de novo* names; and gray, names rejected during the AutoTax workflow.

commonly used classifiers, as revealed by a leave-one-out classification test (see Fig. S3 in the supplemental material).

Since species-level classification is desired wherever possible and the official SILVA taxonomy for bacteria and archaea is not curated at the species level (37), FL-ASVs were also mapped to the 16S rRNA gene sequences from type strains extracted from the SILVA 138 SSURef NR99 database, as they carry official species names. Species-level classifications were assigned to the FL-ASVs if they shared more than 98.7% identity with only one species. If the FL-ASVs matched more than one, they were not classified at the species level due to the high risk of misclassification.

Because the SILVA taxonomy does not provide a complete seven-rank taxonomy for all sequences, the missing classifications are covered with a *de novo* placeholder taxonomy. This taxonomy was created based on the clustering of the FL-ASVs at identity thresholds corresponding to each taxonomic rank (12). The clusters were labeled according to the format denovo_x_y, where x is a one-letter abbreviation for the taxonomic rank (k, p, c, o, f, g, and s), and y represents the number of the FL-ASV, which is the cluster centroid of the particular taxon. Because the applied clustering algorithm processes the sequences sequentially in the order they appear in the input file, sequences are always clustered in the same way, even if additional FL-ASVs are later added to the database. This strategy may not always yield the most optimal clusters, but the reproducibility is critical if the clusters are to be used as a robust placeholder taxonomy.

Merging of the SILVA- and the *de novo*-based taxonomies resulted in a few conflicts, e.g., where different FL-ASVs from the same species associate with more than one genus. In such cases, the genus-level classification for the centroid FL-ASV is adapted

TABLE 2 Numbers and percentages of taxa which were assigned *de novo* placeholder names

Taxa	No. of *de novo* taxa	Percentage
Phylum	1	2.22
Class	12	10.17
Order	66	24.35
Family	276	47.18
Genus	1,324	72.91
Species	3,973	95.28

for all FL-ASVs within that species. However, these types of conflicts only applied to lower rank taxa (species), which were located close to the taxonomic threshold of the higher rank taxa (genus), and it only affected the classification of a low number of FL-ASVs (approximately 1%).

As AutoTax is based on the SILVA taxonomy, the taxonomy generated will change if another version of the SILVA SSURef NR99 reference database is used. Accordingly, users must specify which version of the SILVA database has been used when they publish databases created with AutoTax. We recommend that AutoTax-generated databases are updated when there is a new version of the SILVA database. This ensures that the taxonomy is in agreement with the current central taxonomy.

Taxonomy assignment to FL-ASVs with AutoTax. AutoTax provided placeholder names for many previously undescribed taxa in our FL-ASV database (Table 2, see Fig. S4 in the supplemental material). A total of 95% of all species, 73% of all genera, 47% of all families, and 24% of all orders obtained placeholder names from the *de novo* taxonomy and would otherwise have remained unclassified. The novel taxa were affiliated with several phyla, especially the *Proteobacteria*, *Planctomycetota*, *Patescibacteria*, *Firmicutes*, *Chloroflexi*, *Bacteroidota*, *Actinobacteriota*, and *Acidobacteriota* (Fig. S4). A prominent example is the *Chloroflexi*, where 5 out of 14 orders, 26 out of 34 families, and 142 out of 152 genera observed were assigned a *de novo* placeholder taxonomy. We believe that this method will have important implications for future studies and the accumulation and transfer of knowledge about these taxa in WWTPs, given their high diversity and abundance, and their association with serious operational problems related to the settling of activated sludge (bulking) and foaming (38, 39). It should be noted that the placeholder taxonomy does not provide the same degree of support as traditional phylogenetic analyses and should only be used until an official taxonomy is established.

Improved classification of ASVs from WWTP and anaerobic digesters. To benchmark the FL-ASV database, we classified V1-V3 amplicon data obtained from activated sludge and anaerobic digester samples (Table S2) using this database with the SINTAX classifier and compared the results to classifications obtained using the universal reference databases (Fig. 3a). With the use of the FL-ASV database, many more of the abundant ASVs (>0.01% relative abundance) were classified to the genus and species level (89.9% ± 4.3% and 78.5% ± 4.0%, respectively) than that using popular public reference databases, including SILVA 138 SSURef NR99 (30.4% ± 3.5% and 0%), Greengenes 16S v. 13.5 (24.5% ± 4.4% and 1.4% ± 0.4%), GTDB 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 MiDAS 2.1 (59.7% ± 4.5% and 0.4% ± 0.3%), which is a manually ecosystem-specific curated version of the SILVA 123 SSURef NR99 database.

We have here shown that increased coverage of the rare biosphere can be achieved by adding FL-OTUs clustered at 99% to the FL-ASV database. The FL-OTU-expanded database increased the classification rate at the genus level for the abundant bacteria (>0.01% relative abundance) (94.2% ± 1.4% versus 89.9% ± 4.3%) but reduced classification at the species level (67.7% ± 2.6% versus 78.5% ± 4.0%) (Fig. 3a). When ASVs from the rare biosphere (>0.001% relative abundance) were included in the analysis, the advantage of including FL-OTUs for genus-level classification was even more pronounced (93.0% ± 0.6% versus 80.7% ± 1.7%), and an improved classification rate

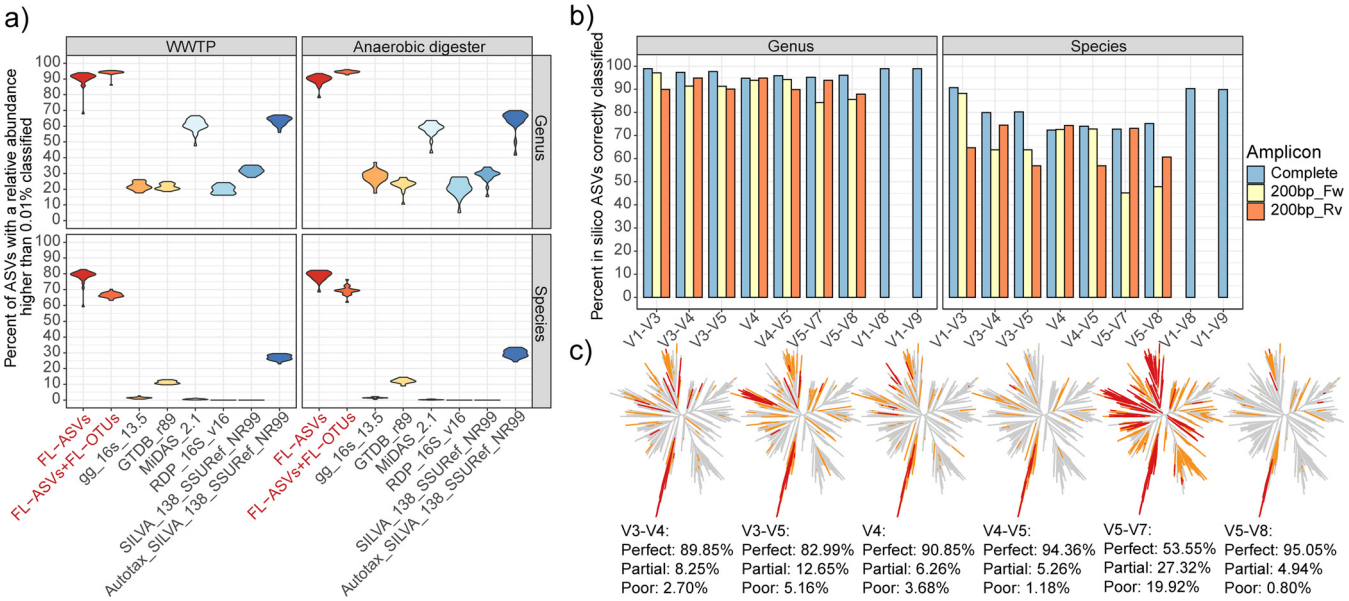

FIG 3 Evaluation of amplicon classification and primer specificity. (a) The fraction of V1-V3 ASVs from activated sludge and anaerobic digester samples with a relative abundance higher than 0.01% that were classified to the genus and species level using the FL-ASV reference database, the FL-OTU expanded 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: V1-V3 (44), V1-V8, V1-V9, V3-V4, and V5-V8 (43); V3-V5 (64); V4 (65); V4-V5 (66); and V5-V7 (67, 68).

at the species level was also observed (69.9% ± 1.5% versus 66.8% ± 2.4%) (see Fig. S5 in the supplemental material). We hypothesize that the reduced classification rate at the species level for the abundant ASVs is caused by sequencing errors as well as low-divergence chimera, which cannot be detected by chimera filtering (40). It should be noted that FL-OTUs are problematic for databases that are to be maintained and updated with additional references in the future. This is because the placeholder taxonomy will change if FL-OTUs are removed, and sequencing errors and chimeras are propagated if they are kept. We therefore recommend that FL-OTUs are added only for exploratory purposes.

To investigate the influence of AutoTax on taxonomic assignment, independent of our ecosystem-specific database, we applied the pipeline directly to high-quality, full-length sequences from the SILVA 138 SSURef NR99 database. This increased the percentage of ASVs classified with SILVA at the genus level from 30.4% ± 3.5% to 63.7% ± 4.8% and at the species-level from 0% to 28.2% ± 2.4%, suggesting that the large universal databases would also benefit from the use of AutoTax. An advantage of the AutoTax-processed SILVA database, which we have made publicly available, is that the placeholder taxonomy is universally applicable, providing a unique opportunity for studying the ecology of unclassified taxa across ecosystems.

Taxonomic resolution of ASVs in combination with a comprehensive reference database. Classification of amplicon sequences can be challenging due to the limited taxonomic information in short-read sequences (12, 13). However, this challenge may change with the access to reference databases with perfect references for the majority of all ASVs and a complete seven-rank taxonomy for all reference sequences. To determine the confidence of the amplicon classification in this scenario, we extracted ASVs *in silico* from the bacterial FL-ASVs corresponding to commonly amplified 16S rRNA regions, including full-length amplicons. These ASVs were classified against the FL-ASV database. We then calculated the fraction of amplicons correctly classified to the same genus and species as their corresponding FL-ASVs (Fig. 3b and Fig. S6a in the supplemental material). Nearly all ASVs (95% to 99%) were assigned to the correct

genus and most (72% to 91%) to the right species, depending on the taxonomic conservation of 16S rRNA region covered by the *in silico* amplicons. The primers targeting the V1-V3 variable region performed exceptionally well for species-level identification (90.7% correct classifications), which is the same as for the full-length 16S rRNA gene amplicons. The commonly used primers targeting the V4 variable region were the worst (72.5% correct classifications). Very few of the sequences that did not receive the same taxonomic classification as their source reference sequence where misclassified ($<$0.2% at genus level and $<$0.8% at species level), with the majority not receiving any classification at the specific taxonomic rank (Fig. S6a).

Sequencing costs on the Illumina platforms can be reduced considerably if single reads are used instead of merged reads. To evaluate the effect of reduced amplicon length, we compared the classification of 200-bp forward reads and reverse reads to those of full-length amplicons (Fig. 3b). The decrease in the percentage of correct classifications was highly dependent on the 16S rRNA region targeted and from which direction the single reads were obtained. For the ecosystem studied here, the V1-V3 and V4-V5 forward read provided almost the same specificity as the full-length amplicons, whereas the reverse reads performed much worse for species-level classification. For the V3-V4, V5-V7, and V5-V8, the reverse reads performed better than the forward reads, revealing the importance of choosing the right direction for single-read amplicons.

To evaluate the effect of sequencing errors and low-divergence chimeras in the reference database, we also classified the *in silico* ASVs against the reference databases expanded with the FL-OTUs (Fig. S6b). The result confirmed our prior observations that the inclusion of error-prone references had a negative impact on our ability to classify short-read amplicons correctly; however, the effect was marginal at the genus level. Full-length amplicons were less affected (Fig. S6b), highlighting a clear advantage of using longer amplicons in combination with universal databases, which are likely to contain sequencing errors and low-divergence chimeras despite chimera filtering (40).

Overall, the analysis demonstrated that confident classification of short-read ASV sequences at the genus to species level is possible. However, it requires a reference database with a complete seven-rank taxonomy and perfect references for the majority of all ASVs. The scripts made available with this study can be used to confirm whether this is the case for samples from other studies.

Ecosystem-specific evaluation of primer bias. When choosing primers for amplicon sequence analyses, it is essential also to take primer bias into account, as some primer sets may result in ecosystem-relevant species being severely underestimated or absent from the analyses (41). The ecosystem-specific FL-ASV databases provide a near-perfect reference to determine the theoretical coverage of different primer sets for the given ecosystem so that an informed selection can be made (42, 43). It should be noted that primers used to generate the full-length 16S rRNA sequences for FL-ASV databases may introduce a bias, and here, we have included primer-free (RNA-based) libraries to account for this. Evaluation of different primer sets using our FL-ASV database revealed a clear taxonomic bias associated with several primer sets (Fig. 3c). The primers targeting the V4-V5 and V5-V8 regions had the best coverage of the FL-ASVs. The V5-V7 primers demonstrated very poor coverage. Because the FL-ASVs were trimmed after the forward priming site of the V1-V3 primers, we were not able to evaluate the coverage of this primer pair here. However, it has previously been shown that the primers have a good overall agreement with metagenomic data for wastewater treatment systems and capture most of the process-critical organisms (44).

Perspectives. High-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax) now allow individual research groups to develop their own FL-ASV ecosystem-specific reference databases for community profiling analyses. In addition, such databases can be used to evaluate the ecosystem-specific coverage and specificity of amplicon primers and fluorescence *in situ* hybridization (FISH) probes. The high quality of the FL-ASVs furthermore allows for the design of new

primers and probes with improved confidence of the coverage and taxonomic resolution when applied within the target ecosystem. Collectively, the approach importantly allows for the identification and subsequent characterization of novel numerically important taxa for the specific environment, which would have otherwise been overlooked.

We acknowledge that the ability to quickly generate new 16S rRNA gene reference databases poses a risk for the development of several competing divergent taxonomies. We, therefore, recommend that the custom databases are only used for exploratory purposes and are combined with traditional phylogenetic analyses of key taxa. Ecosystem-specific databases that are broadly applied, such as the MiDAS database, should be created as open-source community efforts, and universal reference databases processed with AutoTax should be published and maintained in agreement with current developers of such databases. An important benefit of this approach is that the placeholder taxonomy can be used as a common language within the field, or in the case of universal reference databases, across all fields. This has major implications, e.g., in wastewater treatment systems, as it allows for the identification of unclassified taxa that are process-critical and decisive for process performance. If the current universal reference databases are used, the majority of ASVs will not get a genus-level classification, making it impossible to compare their prevalence across studies. Given that hundreds of amplicon-based studies are carried out every year worldwide, a considerable amount of useful information is lost when the data generated across studies are not comparable.

We have chosen to use the SINTAX-classifier for our analyses because it applies a simple algorithm that does not require training, and we expect the results to be less biased. However, some classifiers which use Bayesian inference, e.g., q2-feature-classifier (5), may yield better classifications. Kaehler et al. (45) recently demonstrated that environment-specific taxonomic abundance information could be used as weights for such classifiers to improve the accuracy of the taxonomy assignment. This approach is interesting because the frequency of individual raw full-length 16S rRNA gene sequences used to generate the FL-ASV database may be used as phylogenetically informative weights for the specific ecosystem.

We used SILVA SSURef NR99 as the backbone taxonomy for AutoTax in this study, as it is currently the most comprehensive database. However, we anticipate the GTDB may replace SILVA in a future release of AutoTax when more high-quality genomes and MAGs with 16S rRNA genes (MIMAG standard [11]) are added to the database as the result of advances in long-read sequencing technology (Nanopore and PacBio). This will importantly link the 16S rRNA gene taxonomy with that derived from the more robust phylogenomic-based analyses, creating a unified language across the field of microbiology.

MATERIALS AND METHODS

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

Samples and nucleic acid purification. Activated sludge and anaerobic digester biomass were obtained as frozen aliquots (−80°C) from the MiDAS collection (18). Sample metadata are provided in Table S2. Total nucleic acids were purified from 500 μl of sample thawed on ice using the PowerMicrobiome RNA isolation kit (Mo Bio Laboratories) with the optional phenol-based lysis or with the RiboPure RNA purification kit for bacteria (Thermo Fisher Scientific). Purification was carried out according to the manufacturer recommendations, except that cell lysis was performed in a FastPrep-24 instrument for 4 $\times$ 40 s at 6.0 m/s to increase the yield of nucleic acids from bacteria with sturdy cell walls (41). The samples were incubated on ice for 2 min between each bead beating to minimize heating due to friction. DNA-free total RNA was obtained by treating 41 μl of the purified nucleic acid with the DNase Max kit (Mo Bio Laboratories), followed by clean up using 1.0$\times$ RNAClean XP beads with elution into 25 μl nuclease-free water.

Primer-free full-length 16S rRNA library preparation and sequencing. Purified RNA obtained from biomass samples was pooled separately for each sample source type (activated sludge or anaerobic digester) to give equimolar amounts of 16S rRNA determined based on peak area in the TapeStation

analysis software A.02.02 (SR1). Full-length small subunit (SSU) sequencing libraries were then prepared, as previously described (26). The SSU_rRNA_RT2 (activated sludge) and SSU_rRNA_RT3 (anaerobic digester) reverse transcription primers and the SSU_rRNA_I adaptor were used for the molecular tagging (Table S1), and approximately 1,000,000 tagged molecules from each pooled sample were used to create the clonal library. The final library was sequenced on a HiSeq 2500 instrument using on-board clustering and rapid run mode with a HiSeq paired-end (PE) rapid cluster kit v2 (Illumina) and HiSeq rapid SBS kit v2, 265 cycles (Illumina), as previously described (26). Raw sequence reads were binned based on unique molecular tags, *de novo* assembled into synthetic long-read sequences, and trimmed equivalent to *Escherichia coli* positions 8 and 1507 using the fSSU-pipeline-RNA_v1.2.sh script (https://github.com/KasperSkytte/AutoTax) (26).

Primer-based full-length 16S rRNA gene library preparation and sequencing. The purified nucleic acids obtained from the biomass samples were pooled separately for each sample source type (activated sludge or anaerobic digester) with an equal weight of DNA from each sample. Full-length 16S rRNA sequencing libraries were then prepared, as previously described (26). f16S_rDNA_pcr1_fw1 (activated sludge) or f16S_rDNA_pcr1_fw2 (anaerobic digester) and f16S_rDNA_pcr1_rv (Table S1) were used for the molecular tagging, and approximately 1,000,000 tagged molecules from each pooled sample were used to create the clonal library. The final library was sequenced on a HiSeq 2500 instrument using on-board clustering and rapid run mode with a HiSeq PE rapid cluster kit v2 (Illumina) and HiSeq rapid SBS kit v2, 265 cycles (Illumina), as previously described (26). Raw sequence reads were binned based on unique molecular tags, *de novo* assembled into synthetic long-read sequences, and trimmed equivalent to *E. coli* positions 28 and 1491 using the fSSU-pipeline-DNA_v1.2.sh script (https://github.com/KasperSkytte/AutoTax) (26).

Extraction of high-quality full-length 16S rRNA gene sequences from SILVA. High-quality bacterial and archaeal 16S rRNA gene sequences in the SILVA 138 SSURef NR99 ARB-database were selected using the query pintail_slv = 100 and tax_slv = Bacteria* or tax_slv = Archaea*. Bacterial and archaeal sequences were exported separately in the "fastawide" format after terminal trimming. Bacterial sequences were trimmed between the 27F and 1391R (44) primer binding sites equivalent to positions 1,044 and 41,788 in the global SILVA alignment. Archaeal sequences were trimmed between the 20F (46) and the SSU1000ArR (47) primer binding sites equivalent to positions 1041 and 32818 in the global SILVA alignment. A list of names for full-length sequences spanning the positions above was created using the Extract_full-length_16S_rRNA_names_from_SILVA.sh script, which takes advantage of the fact that ARB uses 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 archaeal sequences without trimming from the SILVA ARB database.

Generation of reference databases using AutoTax. AutoTax was created as a modular multistep Linux BASH script that (i) 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; (ii) 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); and (iii) 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 [48], USEARCH v.11.0.667 [49], SINA v.1.6.0 [50], and R v.3.5.0 with the following packages: biostrings [51], doParallel [52], stringr [53], data.table [54], tidyr [55], and dplyr [56]) into a single BASH script that otherwise requires only 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 is not free and must be purchased online) is available on the GitHub repository online at 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 variants. 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.

Preparation of chimera-filtered full-length 16S rRNA OTUs. Dereplicated sequences (before denoising) from above are clustered at 99% sequence identity using the *usearch -cluster_smallmem* command with the *-id 0.99*, *-maxrejects 0*, *-centroids*, and *-sortedby size* options. Potential chimeras are identified and extracted using the *usearch -uchime2_ref* command with the *-strand plus*, *-mode sensitive*, and *-chimeras* options with the FL-ASVs from above as the reference database. The chimeras are finally removed to create the final FL-OTUs using the *usearch -search_exact* command with the *-strand plus* and *-dbnotmatched* options.

Taxonomy assignment. A complete taxonomy from kingdom to species is automatically assigned to each FL-ASV. In brief, the AutoTax script identifies the closest relative of each FL-ASV in the SILVA database using the *usearch -usearch_global* command, obtains the taxonomy for this sequence, and discards information at taxonomic ranks not supported by the sequence identity and the thresholds for taxonomic ranks proposed by Yarza et al. (12). The identity thresholds used for each of the taxonomic ranks are 75.0%, phylum; 78.5%, class; 82.0%, order; 86.5%, family; 94.5%, genus; and 98.7%, species. For

the species-level classification, the script identifies all type strains within the species-level threshold in the SILVA database and assigns a species-level classification to the FL-ASV if only a single species fits within the threshold. In addition, FL-ASVs are *de novo* clustered using the *usearch -cluster_smallmem* command using the thresholds for each taxonomic rank. The *de novo* clusters are labeled according to the format denovo_x_y, where x is a one-letter abbreviation for the taxonomic rank (k, p, c, o, f, g, and s), and y represents the FL-ASV number of the cluster centroid for each taxon. These labels act as a placeholder taxonomy, where the SILVA taxonomy does not provide any taxonomy information.

Amplicon sequencing and analyses. Bacterial community analyses were performed by amplicon sequencing of the V1-V3 variable region of the 16S rRNA gene as previously described (57) using the 27F (5′-AGAGTTTGATCCTGGCTCAG-3′) (44) and 534R (5′-ATTACCGCGGCTGCTGG-3′) (58) primers. Forward reads were processed using USEARCH v.11.0.667. Raw fastq files were filtered for phiX sequences using *usearch -filter_phix*, trimmed to 250 bp using *usearch -fastx_truncate -trunclen 250*, and quality filtered using *usearch -fastq_filter* with the *-fastq_maxee 1.0* option. The sequences were dereplicated using *usearch -fastx_uniques* with the *-sizeout* option. Exact amplicon sequence variants (ASVs) were resolved using *usearch -unoise3* (4). ASV tables were created by mapping the raw reads to the ASVs using *usearch -otutab* with the *-zotus* and *-strand both* options. Taxonomy was assigned to ASVs using *usearch -sintax* with *-strand both* and *-sintax_cutoff 0.8* (13).

Construction of phylogenetic trees and primer evaluation. FL-ASVs aligned to the SILVA 138 NR99 ARB database were obtained from the AutoTax output (temp/FL-ASVs_SILVA_aligned.fa) and loaded into the SILVA 138 NR99 ARB database. All bacterial FL-ASVs were selected and exported as a FASTA file using the ssuref:bacteria positional variability by parsimony filter. A rough tree was created from the alignment using FastTree v.2.1.10 (59) with the *-nt -gtr -gamma* options and loaded into ARB. The specificity of commonly used amplicon primers was determined for each FL-ASV using the *analyze_primers.py* script from PrimerProspector v.1.0.1 (42). The specificity of primer sets was defined based on the overall weighted scores (OWSs) for the primer with the highest score as follows: perfect hit (OWS, 0), partial hit (OWS, >0 and ≤1), and poor hit (OWS, >1). A comma-separated table with the specificity of each primer set for each FL-ASV was made in R, loaded into ARB, and used to color the tree.

Data analyses and visualization. USEARCH v.11.0.667 was used for mapping sequences to references with *-usearch_global -id 0 -maxrejects 0 -maxaccepts 0 -top_hit_only -strand plus*, unless otherwise stated. Data were imported into R v.3.6.3 (60) using RStudio IDE (61), aggregated using the tidyverse package v.1.2.1 (https://www.tidyverse.org/), and analyzed and visualized using ggplot2 v.3.1.0 (62) and ampvis2 v.2.4.0 (63).

Data availability. Raw and assembled sequencing data are available at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under the project number PRJEB26558. The AutoTax script is available online at https://github.com/KasperSkytte/AutoTax. The AutoTax-processed FL-ASV and FL-OTU expanded reference database in SINTAX and QIIME formats is available online at https://figshare.com/articles/Data_used_in_AutoTax_paper/12377741/1. The AutoTax-processed SILVA 138 SSURef NR99 database in SINTAX, and QIIME formats is available online at https://doi.org/10.6084/m9.figshare.12366626. R-markdown scripts used for data analyses and figures are available online at https://github.com/msdueholm/Publications/tree/master/Dueholm2020a.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.04 MB.

FIG S1, PDF file, 0.3 MB.

FIG S2, PDF file, 0.2 MB.

FIG S3, PDF file, 0.2 MB.

FIG S4, PDF file, 0.1 MB.

FIG S5, PDF file, 0.7 MB.

FIG S6, PDF file, 0.2 MB.

TABLE S1, DOCX file, 0.01 MB.

TABLE S2, DOCX file, 0.02 MB.

TABLE S3, DOCX file, 0.01 MB.

ACKNOWLEDGMENTS

We thank the 22 wastewater treatment plants involved in the project for providing samples. This work was supported by the Danish Research Council (6111-00617A to P.H.N.) and the Villum Foundation (13351 and 16578 to P.H.N.). We declare no conflict of interest.

REFERENCES

1. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: a phylogenetic perspective. Science 350:aac9323. https://doi.org/10.1126/science.aac9323.
2. Amir A, Daniel M, Navas-Molina J, Kopylova E, Morton J, Xu ZZ, Eric K, Thompson L, Hyde E, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191-16. https://doi.org/10.1128/mSystems.00191-16.
3. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP.

2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869.
4. Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv https://doi.org/10.1101/081257.
5. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90. https://doi.org/10.1186/s40168-018-0470-z.
6. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267. https://doi.org/10.1128/AEM.00062-07.
7. Edgar R. 2016. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv https://doi.org/10.1101/074161.
8. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639–2643. https://doi.org/10.1038/ismej.2017.119.
9. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, 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, The Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.
10. Mcdonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, Goldasich D, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, The American Gut Consortium, Knight R. 2018. American gut: an open platform for citizen science. mSystems 3:e00031-18. https://doi.org/10.1128/mSystems.00031-18.
11. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, The Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893.
12. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645. https://doi.org/10.1038/nrmicro3330.
13. Edgar RC. 2018. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 6:e4652. https://doi.org/10.7717/peerj.4652.
14. Desantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072. https://doi.org/10.1128/AEM.03006-05.
15. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and Web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219.
16. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642. https://doi.org/10.1093/nar/gkt1244.
17. Louca S, Mazel F, Doebeli M, Parfrey LW. 2019. A census-based estimate of Earth's bacterial and archaeal diversity. PLoS Biol 17:e3000106. https://doi.org/10.1371/journal.pbio.3000106.
18. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM, Albertsen M, Nielsen PH. 2017. MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. Database 2017:bax016. https://doi.org/10.1093/database/bax016.
19. Mikaelyan A, Köhler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A. 2015. Classifying the bacterial gut microbiota of termites and cockroaches: a curated phylogenetic reference database (DictDb). Syst Appl Microbiol 38:472–482. https://doi.org/10.1016/j.syapm.2015.07.004.
20. Ritari J, Salojärvi J, Lahti L, de Vos WM. 2015. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 16:1056. https://doi.org/10.1186/s12864-015-2265-y.
21. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. 2010. The Human Oral Microbiome Database: a Web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford) 2010:baq013. https://doi.org/10.1093/database/baq013.
22. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. 2011. A guide to the natural history of freshwater lake bacteria. Microbiol Mol Biol Rev 75:14–49. https://doi.org/10.1128/MMBR.00028-10.
23. Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD. 2018. TaxAss: leveraging a custom freshwater database achieves fine-scale taxonomic resolution. mSphere 3:e00327-18. https://doi.org/10.1128/mSphere.00327-18.
24. Newton ILG, Roeselers G. 2012. The effect of training set on the classification of honey bee gut microbiota using the Naïve Bayesian Classifier. BMC Microbiol 12:221. https://doi.org/10.1186/1471-2180-12-221.
25. Seedorf H, Kittelmann S, Henderson G, Janssen PH. 2014. RIM-DB: a taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. PeerJ 2:e494. https://doi.org/10.7717/peerj.494.
26. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. 2018. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol 36:190–195. https://doi.org/10.1038/nbt.4045.
27. Burke CM, Darling AE. 2016. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. PeerJ 4:e2492. https://doi.org/10.7717/peerj.2492.
28. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 47:e103. https://doi.org/10.1093/nar/gkz569.
29. Karst SM, Ziels RM, Kirkegaard RH, Albertsen M. 2019. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing. bioRxiv https://doi.org/10.1101/645903.
30. Wu L, Ning D, Zhang B, Li Y, Zhang P, Shan X, Zhang Q, Brown M, Li Z, Van Nostrand JD, Ling F, Xiao N, Zhang Y, Vierheilig J, Wells GF, Yang Y, Deng Y, Tu Q, Wang A, Global Water Microbiome Consortium, Zhang T, He Z, Keller J, Nielsen PH, Alvarez PJJ, Criddle CS, Wagner M, Tiedje JM, He Q, Curtis TP, Stahl DA, Alvarez-Cohen L, Rittmann BE, Wen X, Zhou J. 2019. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat Microbiol 4:1183–1195. https://doi.org/10.1038/s41564-019-0426-5.
31. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229.
32. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
33. Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 38:701–707. https://doi.org/10.1038/s41587-020-0422-6.
34. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M. 2020. Connecting structure to function with the recovery of over 1000 high-quality activated sludge metagenome-assembled genomes encoding full-length rRNA genes using long-read sequencing. bioRxiv https://doi.org/10.1101/2020.05.12.088096.
35. Isazadeh S, Jauffur S, Frigon D. 2016. Bacterial community assembly in activated sludge: mapping beta diversity across environmental variables. MicrobiologyOpen 5:1050–1060. https://doi.org/10.1002/mbo3.388.
36. Gonzalez-Martinez A, Rodriguez-Sanchez A, Lotti T, Garcia-Ruiz MJ, Oso-

rio F, Gonzalez-Lopez J, Van Loosdrecht MCM. 2016. Comparison of bacterial communities of conventional and A-stage activated sludge systems. Sci Rep 6:18786. https://doi.org/10.1038/srep18786.
37. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 42:D643–D648. https://doi.org/10.1093/nar/gkt1209.
38. McIlroy SJ, Karst SM, Nierychlo M, Dueholm MS, Albertsen M, Kirkegaard RH, Seviour RJ, Nielsen PH. 2016. Genomic and in situ investigations of the novel uncultured Chloroflexi associated with 0092 morphotype filamentous bulking in activated sludge. ISME J 10:2223–2234. https://doi.org/10.1038/ismej.2016.14.
39. Nierychlo M, Mcilroy SJ, Kucheryavskiy S, Jiang C, Ziegler AS, Kondrotaite Z, Stokholm-Bjerregaard M, Nielsen PH. 2020. Candidatus Amarolinea and Candidatus Microthrix are mainly responsible for filamentous bulking in municipal Danish wastewater treatment plants. Front Microbiol 11:1214. https://doi.org/10.3389/fmicb.2020.01214.
40. Edgar RC. 2016. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv https://doi.org/10.1101/074252.
41. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. 2015. Back to basics—the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. PLoS One 10: e0132783. https://doi.org/10.1371/journal.pone.0132783.
42. Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. 2011. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics 27: 1159–1161. https://doi.org/10.1093/bioinformatics/btr087.
43. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41:e1. https://doi.org/10.1093/nar/gks808.
44. Lane DJ. 1991. 16S/23S rRNA sequencing, p 115–175. In Stackebrandt E, Goodfellow M (ed), Nucleic acid techniques in bacterial systematics. John Wiley and Sons, Chichester, United Kingdom.
45. Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA. 2019. Species abundance information improves sequence taxonomy classification accuracy. Nat Commun 10:4643. https://doi.org/10.1038/s41467-019-12669-6.
46. Massana R, Murray AE, Preston CM, DeLong EF. 1997. Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. Appl Environ Microbiol 63:50–56. https://doi.org/10.1128/AEM.63.1.50-56.1997.
47. Bahram M, Anslan S, Hildebrand F, Bork P, Tedersoo L. 2019. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. Environ Microbiol Rep 11:487–494. https://doi.org/10.1111/1758-2229.12684.
48. Tange O. 2018. GNU Parallel 2018. Ole Tange https://doi.org/10.5281/zenodo.1146014.
49. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
50. Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829. https://doi.org/10.1093/bioinformatics/bts252.
51. Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2019. Biostrings: efficient manipulation of biological strings.
52. Microsoft Corporation, Weston S. 2019. doParallel: Foreach Parallel Adaptor for the "parallel" package.
53. Wickham H. 2019. stringr: simple, consistent wrappers for common string operations.
54. Dowle M, Srinivasan A. 2019. data.table: extension of "data.frame."
55. Wickham H, Henry L. 2019. tidyr: easily tidy data with "spread()" and "gather()" functions.
56. Wickham H, François R, Henry L, Müller K. 2019. dplyr: a grammar of data manipulation.
57. Kirkegaard RH, McIlroy SJ, Kristensen JM, Nierychlo M, Karst SM, Dueholm MS, Albertsen M, Nielsen PH. 2017. The impact of immigration on microbial community composition in full-scale anaerobic digesters. Sci Rep 7:9343. https://doi.org/10.1038/s41598-017-09303-0.
58. Muyzer G, de Waal EC, Uitterlinden AG. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol 59:695–700. https://doi.org/10.1128/AEM.59.3.695-700.1993.
59. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
60. R Core Team. 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
61. RStudio Team. 2015. RStudio: integrated development environment for R. RStudio, PBC, Boston, MA.
62. Wickham H. 2009. ggplot2—elegant graphics for data analysis. Springer Science & Business Media, New York, NY.
63. Andersen KS, Kirkegaard RH, Karst SM, Albertsen M. 2018. ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. bioRxiv https://doi.org/10.1101/299537.
64. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal K, Baker CC, Francesco VDi, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. 2009. The NIH Human Microbiome Project. Genome Res 19:2317–2323. https://doi.org/10.1101/gr.096651.109.
65. Apprill A, Mcnally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol 75:129–137. https://doi.org/10.3354/ame01753.
66. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414. https://doi.org/10.1111/1462-2920.13023.
67. Bodenhausen N, Horton MW, Bergelson J. 2013. Bacterial communities associated with the leaves and the roots of Arabidopsis thaliana. PLoS One 8:e56329. https://doi.org/10.1371/journal.pone.0056329.
68. Chelius MK, Triplett EW. 2001. The diversity of archaea and bacteria in association with the roots of Zea mays L. Microb Ecol 41:252–263. https://doi.org/10.1007/s002480000087.
69. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet J 17:10–12. https://doi.org/10.14806/ej.17.1.200.