MiDAS 5: Global diversity of bacteria and archaea in anaerobic digesters

Authors: Morten Kam Dahl Dueholm[1,*], Kasper Skytte Andersen[1], Anne-Kirstine C. Petersen[1], Vibeke Rudkjøbing[1], Madalena Alves[2], Yadira Bajón-Fernández[3], Damien Batstone[4], Caitlyn Butler[5], Mercedes Cecilia Cruz[6], Åsa Davidsson[7], Leonardo Erijman[8], Christof Holliger[9], Konrad Koch[10], Norbert Kreuzinger[11], Changsoo Lee[12], Gerasimos Lyberatos[13], Srikanth Mutnuri[14], Vincent O'Flaherty[15], Piotr Oleskowicz-Popiel[16], Dana Pokorna[17], Veronica Rajal[6], Michael Recktenwald[18], Jorge Rodríguez[19], Pascal E. Saikaly[20], Nick Tooker[5], Julia Vierheilig[11], Jo De Vrieze[21], Christian Wurzbacher[10] & Per Halkjær Nielsen[1,*]

Affiliations:

[1]Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark. [2]Centre of Biological Engineering, University of Minho, Portugal. [3]School of Water, Energy and Environment, Cranfield University. Cranfield, United Kingdom. [4]Australian Centre for Water and Environmental Biotechnology (ACWEB), The University of Queensland, Australia. [5]Department of Civil and Environmental Engineering, University of Massachusetts Amherst, MA, USA. [6]National University of Salta, Salta, Argentina. [7]Department of Chemical Engineering, Lund University, Sweden. [8]INGEBI-CONICET, University of Buenos Aires, Argentina. [9]Laboratory for Environmental Biotechnology, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. [10]Chair of Urban Water Systems Engineering, Technical University of Munich (TUM), Garching, Germany. [11]Institute of Water Quality and Resource Management, TU Wien, Austria. [12]Department of Urban and Environmental Engineering & Graduate School of Carbon Neutrality, Ulsan National Institute of Science and Technology (UNIST), South Korea. [13]School of Chemical Engineering, National Technical University of Athens, Greece. [14]Applied Environmental Biotechnology Laboratory, Birla Institute of Technology and Science (BITS-Pilani), India. [15]School of Biological and Chemical Sciences and Ryan Institute, University of Galway, Ireland. [16]Water Supply and Bioeconomy Division, Faculty of Environmental Engineering and Energy, Poznan University of Technology, Poland. [17]Department of Water Technology and Environmental Engineering, University of Chemistry and Technology Prague, Czech Republic. [18]Research Scientist at Kemira Oyj, Espoo R&D Center, Finland. [19]Chemical Engineering Department, Khalifa University, United Arab Emirates. [20]Environmental Science and Engineering Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. [21]Center for Microbial Ecology and Technology (CMET), Ghent University, Belgium.

*Correspondence to: Per Halkjær Nielsen, Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220 Aalborg, Denmark; Phone: +45 2173 5089; Fax: Not available; E-mail: phn@bio.aau.dk or Morten Kam Dahl Dueholm, Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220 Aalborg, Denmark; Phone: (+45) 9940 8508; Fax: Not available; E-mail: md@bio.aau.dk

Running title: Global microbiota of anaerobic digesters

Abstract

Anaerobic digestion represents a key biotechnology for the transformation of organic waste into renewable energy (biogas) and relies on complex microbial communities that work in concert to degrade the complex substrates into methane and carbon dioxide. Here, we sequenced more than half a million high-quality, full-length 16S rRNA gene sequences from 285 full-scale anaerobic digesters (ADs) across the world to expand our knowledge about diversity and function of the bacteria and archaea in ADs. The sequences were processed into full-length 16S rRNA amplicon sequence variants (FL-ASVs), which were added to the MiDAS 4 database for bacteria and archaea in wastewater treatment systems to create MiDAS 5. The expansion of the MiDAS database significantly increased the coverage for bacteria and archaea in ADs worldwide, leading to an improved rate of genus and species-level classification. Using MiDAS 5, we carried out an amplicon-based, global-scale microbial community profiling of the sampled ADs using three common sets of primers targeting different regions of the 16S rRNA gene in bacteria and/or archaea. We revealed how environmental conditions and biogeography shape the AD microbiota. We also identify core and conditionally rare or abundant taxa, encompassing 692 genera and 1013 species. These represent 84-99% and 18-61% of the accumulated read abundance respectively, across samples depending on the amplicon primers used. Finally, we examined the global diversity of functional groups with known importance for the anaerobic digestion process. Our online global MiDAS Field Guide presents the data generated in this study and summarizes present knowledge about all taxa.

Introduction:

Anaerobic digestion has gained attention as an important, sustainable biotechnology as it provides several benefits that align with the goals of sustainability. It can help to produce renewable energy (biogas) from organic waste such as manure, food waste, and sludge from wastewater treatment plants (WWTPs) [1,2]. The anaerobic digestion process also reduces pathogens and the amount of organic waste that is sent to landfills, thereby reducing methane emissions and supporting sustainable waste management practices [1]. Finally, the fertilizer that is produced as a byproduct of anaerobic digestion can be used to support sustainable agriculture, reducing the need for synthetic fertilizers that can have negative environmental impacts [3,4].

The anaerobic digestion process relies on the microbial degradation and conversion of organic matter, which requires a complex interplay between several functional guilds. These include hydrolyzing, acidogenic, and acetogenic syntrophic bacteria as well as methanogenic archaea [5]. The taxonomy is poorly characterized for many of the microorganisms in anaerobic digesters (ADs), and even among the most abundant taxa many lack genus- or species-level classifications [6]. To optimize performance, a comprehensive knowledge about microbial immigration/competition, environmental/operational conditions, and taxonomy is essential [7–9]. Recent microbial surveys have increased our knowledge about the anaerobic digestion process [7,10–16]. However, sharing knowledge across studies is still hindered by the absence of standardized protocols and a common reference database with a unifying taxonomy [17,18]. To facilitate collaboration and knowledge sharing, it is essential to establish these standard protocols and resources.

The Microbial Database for Activated Sludge and Anaerobic Digesters (MiDAS) project was established as an open-source platform for sharing updated knowledge about the physiology and ecology of the important microorganisms present in engineered ecosystems of activated sludge plants, ADs, and related WWTPs [17–20]. MiDAS provides standardized protocols for microbial profiling of microbes in wastewater treatment systems [21], an ecosystem-specific full-length 16S rRNA gene reference database [18,20], and a field guide where knowledge about the specific genera are stored and shared (https://www.midasfieldguide.org).

The MiDAS 16S rRNA gene reference database was created based on millions of high-quality, chimera-free, full-length 16S rRNA genes resolved into amplicon sequence variants (ASVs) and classified using automated taxonomy assignment (AutoTax) [6,18,20]. AutoTax provides a comprehensive seven-rank taxonomy (kingdom to species-level) for all reference sequences based on the most recent version of the SILVA SSURef 99 NR taxonomy and includes a robust placeholder taxonomy for lineages without an official taxonomy [6]. The placeholder taxa are easily distinguishable by their names, formatted as 'midas_x_y', where 'x' indicates the taxonomic rank and 'y' is a numerical

identifier. This naming convention facilitates the study of unclassified alongside classified taxa across various taxonomic ranks. The placeholder taxonomy should not be seen as a replacement for proper taxonomic classifications but can pinpoint important lineages that should be studied in depth using phylogenomics [22–26].

The MiDAS 16S rRNA gene reference database (MiDAS 4.8.1) currently contains reference sequences from anaerobic digesters located at WWTPs in Denmark and WWTPs worldwide [20]. However, it may not provide comprehensive coverage for all important microbes found in ADs treating other types of waste or in other locations.

In this study, we introduce MiDAS 5, an updated version of MiDAS 4 expanded with more than half a million high-quality, full-length archaeal and bacterial 16S rRNA gene sequences from 285 ADs worldwide treating different types of biowaste. We carried out a global survey of ADs using three commonly used short-read amplicon primer sets targeting bacteria (V1-V3), archaea (V3-V5), and both (V4). This data was used in combination with MiDAS 5 to i) link the global diversity of bacteria and archaea to biogeography and environmental factors, ii) identify important core taxa, and iii) uncover the global diversity within selected functional guilds. The results provide a solid foundation for future research on AD microbiology.

Results and Discussion:

The MiDAS Global Consortium for Anaerobic Digesters was established in 2018 to coordinate the sampling and collection of metadata from ADs worldwide (Supplementary Data 1). Samples were obtained in duplicates from 285 ADs in 196 cities in 19 countries on five continents (Fig. 1a). Most of the ADs treated surplus sludge from wastewater treatment plants (69.8%) (Fig. 1b). However, ADs treating food waste (8.1%), industrial waste (7.4%), and manure (5.3%) were also included in the survey. Most of the ADs were mesophilic (86.0%), few were thermophilic (6.0%), and some missed temperature data (8.1%). The main digester technology used was completely mixed reactors (67.7%) followed by two-stage reactors (12.6%). A few upflow anaerobic sludge blanket (UASB) and other types were also sampled to expand the diversity of digester types.

Expanding the MiDAS database with reference from global ADs

To expand the MiDAS database with sequences from ADs across the globe, we applied high-fidelity, full-length 16S rRNA gene sequencing on all samples collected in this study. More than half a million full-length 16S rRNA gene sequence reads, representing both bacteria and archaea, were obtained after quality filtering and primer trimming. After processing the sequence reads with AutoTax to produce full-length 16S rRNA gene ASVs (FL-ASVs), these were compared and added to the existing 90,164 FL-ASVs in the MiDAS 4.8.1 database. The combined number was then deduplicated, resulting in a total of 120,408 non-redundant FL-ASV reference sequences in the

expanded MiDAS 5 database. This represents an increase of 30,246 new FL-ASVs when compared to the previous version.

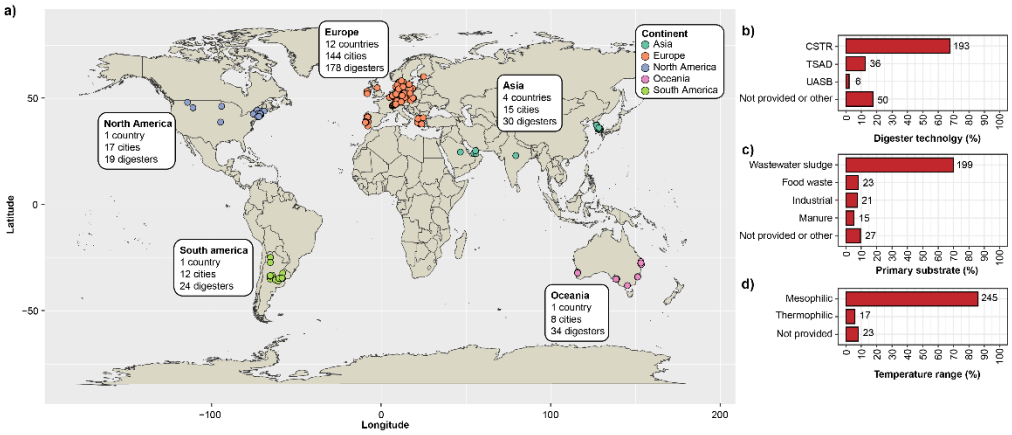

Fig. 1: Sampling of anaerobic digesters (ADs) across the world. a) Geographical distribution of ADs included. b) Distribution of digester technologies. CSTR: Continuous stirred-tank reactor; TSAD: Two-stage anaerobic digestion; UASB: Upflow anaerobic sludge blanket. c) Distribution of primary substrates. d) Distribution of digester temperatures. The values next to the bars are the number of ADs in each group.

Table 1: Sequence novelty of FL-ASVs obtained in this study. Sequence novelty was determined based on the percentage identity between each of the 30,246 new FL-ASV and their closest relative in the databases indicated and identity thresholds for each taxonomic rank proposed by Yarza *et al.* [27] shown in the parentheses.

	SILVA 138.1 SSURef NR99		MiDAS 4.8.1	
	Sequences	Percentage	Sequences	Percentage
Phylum (<75.0%)	28	0.09%	67	0.22%
Class (<78.5%)	55	0.18%	186	0.61%
Order (<82.0%)	112	0.37%	462	1.53%
Family (<86.5%)	354	1.17%	1,483	4.90%
Genus (<94.5%)	5,240	17.32%	9,419	31.14%
Species (<98.7%)	15,806	52.26%	16,863	55.75%

The novelty of the 30,246 new FL-ASVs were determined based on the percent identity shared with their closest relatives in the SILVA 138.1 SSURef NR99 and MiDAS 4.8.1 database using the threshold for each taxonomic rank proposed by Yarza *et al.* [27] (Table 1). It should be noted that these thresholds do not uniformly apply across the bacterial phylogenetic tree; therefore, our taxonomic assignments should be considered as approximations intended to facilitate biological interpretation. 17% and 31% of the new FL-ASV lacked genus-level homologs (≥94.5% identity) and 52% and 56% were without species-level homologs (≥98.7% identity) in SILVA 138.1 and MiDAS 4, respectively. This suggests a substantial increase in the diversity within the MiDAS 5 database.

181 ***MiDAS 5 introduces many new taxa***

182 To investigate how the new FL-ASVs affected the taxonomic diversity in the MiDAS
183 database, we determined the number of additional taxa introduced at different
184 taxonomic ranks (Table 2). A substantial increase in diversity was observed with the
185 addition of 2,770 new genera (29.2% increase) and 8,858 new species (28.3% increase).
186 However, many additional taxa were also introduced at higher taxonomic ranks
187 including six more bacterial and five more archaeal phyla previously known from the
188 SILVA taxonomy. In addition, we identified nine lineages classified as MiDAS
189 placeholder phyla. However, phylogenetic analysis revealed that these lineages branch
190 closely to mitochondrial sequences, indicating they are likely mitochondrial in origin.
191 The largest percentage of the new FL-ASVs (42.8%) were found within the Firmicutes
192 (Supplementary Fig. 1a). Firmicutes often occur in high abundance in ADs, where they
193 are involved in fermentation and thereby directly stimulate biogas yields [7,10,13,15,28]. A
194 closer look into the expanded diversity within the Firmicutes revealed that new FL-
195 ASVs were associated with several families (Supplementary Fig. 1b), including
196 Hungateiclostridiaceae (1,324 FL-ASVs), Lachnospiraceae (788 FL-ASVs),
197 Peptostreptococcales-Tissierellales Family_XI (763 FL-ASVs), Christensenellaceae
198 (754 FL-ASVs), Caldicoprobacteraceae (620 FL-ASVs), and Syntrophomonadaceae
199 (555 FL-ASVs). The Syntrophomonadaceae are of special relevance, as this family
200 includes several syntrophic fatty acid degrading bacteria, which are often the metabolic
201 bottleneck in the overall ADs process [29,30].

202 Table 2: New taxa introduced with MiDAS 5. The number of new taxa represent unique taxa at the
203 different taxonomic ranks that were not part of MiDAS 4.8.1 [20] and includes both official taxonomic
204 names and de novo placeholder names provided by AutoTax [6].

	Total taxa	New taxa	Increase (%)
Phylum	105	20	23.5%
Class	259	37	16.7%
Order	727	104	16.7%
Family	2,212	360	19.4%
Genus	12,254	2,770	29.2%
Species	40,207	8,858	28.3%

205

206 ***MiDAS 5 provides improved coverage and classifications for AD microbiota***
207 The performance of the MiDAS 5 database was evaluated based on three ASV-
208 resolved, short-read, 16S rRNA gene amplicon datasets generated from the AD samples
209 collected in this study (Fig. 2). The V1-V3 amplicons include only bacteria and provide
210 high phylogenetic resolution. However, the primers targeting this region have a lower
211 coverage for the known bacterial diversity according to in silico evaluations [6,31]. The
212 V4 amplicons include both bacteria and some archaeal lineages and are commonly used
213 due to a very good coverage of the known bacterial diversity. However, the amplicons
214 have a weaker phylogenetic resolution compared to V1-V3, which in many cases
215 prevent species-level classifications [6,31]. The V3-V5 amplicons cover mainly archaea
216 and have previously been used to describe their diversity in ADs [7,10].
217
218 Our initial analysis involved non-heuristic mapping of short-read ASVs against MiDAS
219 5 and other widely used reference databases, including the newly released GreenGenes2
220 [32]. This step allowed us to establish the percent identity between each ASV and its
221 closest match across the databases. We then calculated the percentage of ASVs that
222 have high-identity matches (≥99% identity) in each sample and database. To focus on
223 active microbial populations, we excluded ASVs representing the rare biosphere (those
224 with <0.01% relative abundance), which are often enriched in non-growing organisms
225 and environmental DNA [7,10]. MiDAS 5 performed exceptionally well for bacteria with
226 high-identity hits of 94.8%±4.2% (mean±SD) for V1-V3 and 96.3%±2.1% for V4
227 ASVs, compared to 67.9%±19.7% and 71.4%±16.1% for MiDAS 4, and 61.1%±9.2%
228 and 77.1%±7.8% for SILVA v.138.1 (Fig. 2). The complete GreenGenes2 database
229 displayed a coverage close to that of MiDAS 5 for V4 ASVs (95.4%±3.3%) but a much
230 lower coverage for V1-V3 (32.1%±8.9%). The reason is that the complete
231 GreenGenes2 database contains V4 ASVs from Qiita [33] in addition to full-length 16S
232 rRNA gene sequences [32]. For the V3-V5 archaeal dataset, an increase in coverage was
233 observed from 33.5%±7.0% with MiDAS 4 to 55.9%±9.5% with MiDAS 5. However,
234 the SILVA database (67.0%±11.0%) and the complete GTDB database (69.2%±13.0%)
235 provide even better coverage. The lower coverage for archaea compared to bacteria in
236 MiDAS 5 is likely due to reduced sequencing efforts and the challenges in designing
237 effective universal primers for archaeal full-length 16S rRNA gene sequencing [34,35].
238
239 Because the sampling of ADs was directed towards mesophilic digesters treating
240 surplus sludge from WWTPs, we also evaluated the MiDAS 5 coverage for ADs
241 treating different primary substrates and temperatures (Supplementary Fig. 2). MiDAS
242 5 gave very good coverage for all sample types supporting the general applicability of
243 the reference database for ADs. Finally, to provide additional support for the general
244 applicability of the MiDAS 5 database, we evaluated it based on previously published
245 V4-V5 amplicon data from 90 full-scale ADs at 51 municipal WWTPs unrelated to this
246 study [14]. MiDAS 5 contained high-identity hits for 91.8%±6.8% of the ASVs
247 (Supplementary Fig. 3).

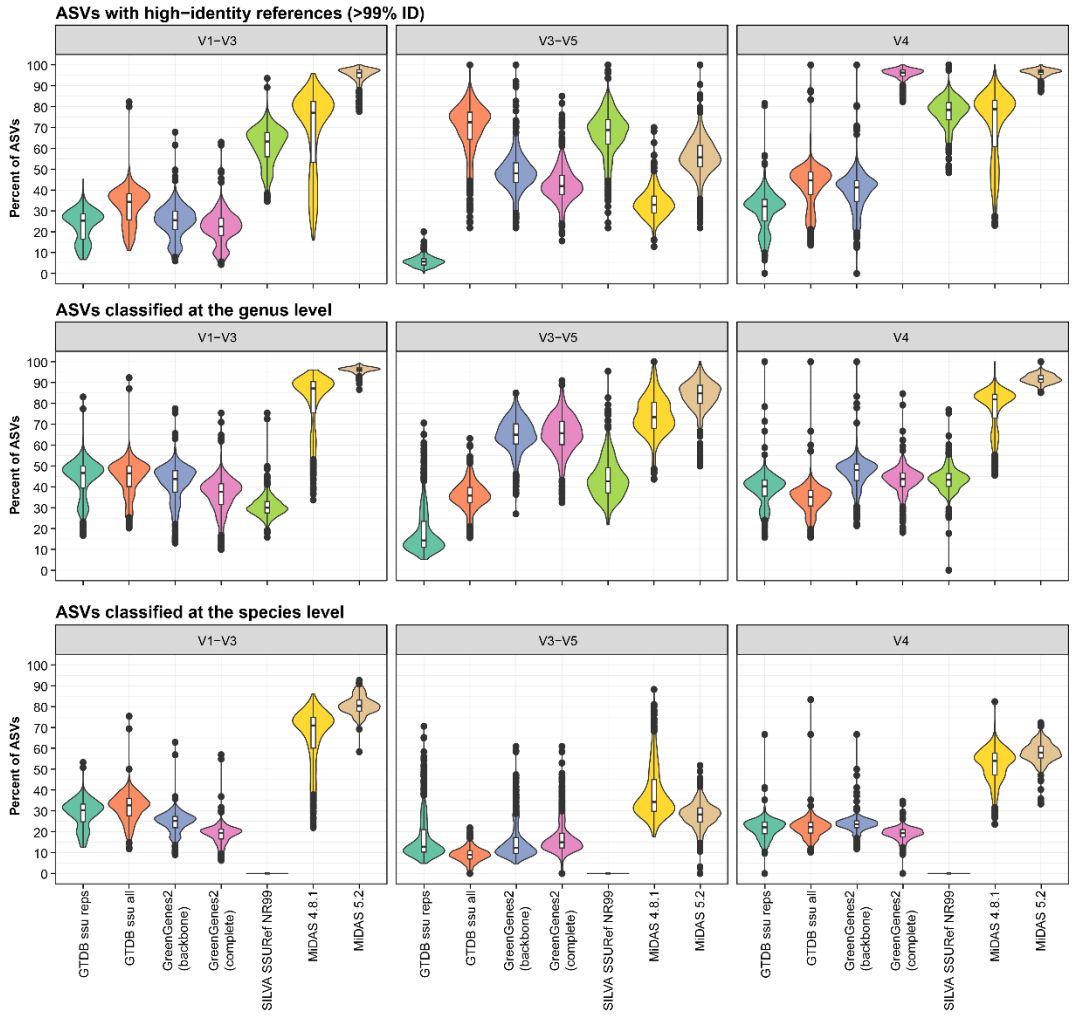

Fig. 2: Database evaluation based on short-amplicon data from this study. The ASVs for each of the samples were filtered based on their relative abundance (only ASVs with ≥0.01% relative abundance were kept) before the analyses. The percentage of the microbial community represented by the remaining ASVs after the filtering was 95.44% ± 2.23% (mean±SD) for V1-V3 amplicons (only bacteria), 99.65% ± 0.17% for V3-V5 amplicons (mainly archaea), and 97.34% ± 2.01% for V4 amplicons (bacteria and archaea) across samples. High-identity (≥99%) hits were determined by stringent mapping of ASVs to each reference database. Classification of ASVs was done using the SINTAX classifier. The violin and box plots illustrate the distribution of the percentage of ASVs with high-identity hits or genus/species-level classifications for each database, analyzed across 570 biologically independent samples, including two biological replicates for each digester. Box plots indicate median (middle line), 25th, 75th percentile (box), and the min and max values after removing outliers based on 1.5x interquartile range (whiskers). Outliers have been removed from the box plots to ease visualization. Different colors are used to distinguish the different databases: GTDB_bac120_ssu_reps_r214, GTDB_ssu_all_r214, GreenGenes2_2022_10 (backbone and complete database), SILVA 138.1 SSURef NR99, MiDAS 4.8.1, and MiDAS 5.2.

Our second database evaluation was based on the classification of ASVs from each amplicon dataset using the SINTAX classifier (Fig. 2). We found that MiDAS 5 greatly improved the rates of genus-level classification (96.3%±1.4% for V1-V3, 91.5%±2.6% for V4, and 82.6%±7.5% for V3-V5) compared to MiDAS 4 (80.2%±14.9% for V1-V3, 77.3%±10.5% for V4, and 74.7%±9.3% for V3-V5), and the rates of classification were more than two fold higher than those obtained with any of the other evaluated databases for bacteria and also higher for archaea. Analysis of species-level classifications revealed similar improvements with MiDAS 5 for bacteria (Fig. 2). However, a decrease in species-level classifications was observed between MiDAS 4 and 5 for the archaeal V3-V5 dataset. We hypothesize that this effect relates to over-classifications with MiDAS 4 due to the lack of appropriate reference sequences in MiDAS 4.

Finally, we investigated if the additional reference sequences introduced in MiDAS 5 could improve classification of amplicon data from WWTPs based on data from the MiDAS global sampling of WWTPs [20] and the Global Water Microbiome Consortium project [36] (Supplementary Fig. 4). Interestingly, no statistically significant improvements were observed. This highlights that most of the added references originated from anaerobic digester-specific taxa.

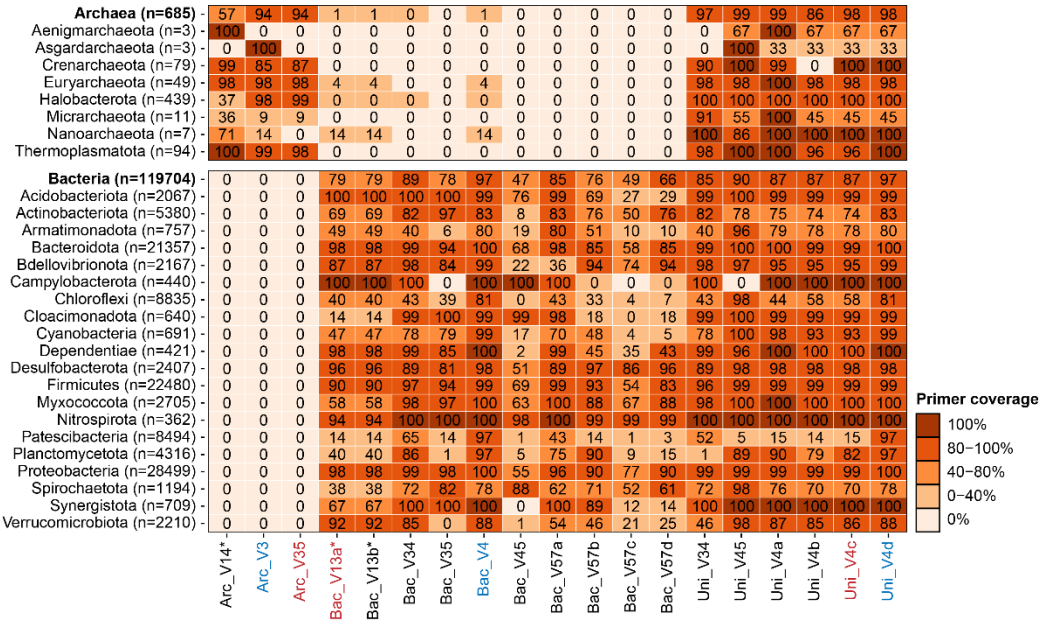

Fig. 3: Ecosystem-specific primer coverage for all archaea and bacteria and the 20 most diverse bacterial and all archaeal phyla based on unique FL-ASVs. The number of FL-ASVs for each taxon (n) is provided next to the names. The coverage was determined as the percentage of FL-ASVs in the MiDAS 5.2 database with perfect hits for both forward and reverse primers. The primer pairs marked in red were used in the current study and the ones marked in blue are those recommended here based on coverage. Detailed information of all primer pairs and coverage information for all taxa in MiDAS 5.2 are provided in Supplementary Data 2. *Only the reverse primer was evaluated for these primer pairs because the forward primer was used to create the reference sequences in MiDAS. The coverage might therefore be overestimated for these primer pairs.

294 ***Evaluation of 16S rRNA gene amplicon primers for community profiling of ADs***
295 The comprehensive ASV-resolved MiDAS 5 database provides a unique opportunity
296 to determine the theoretical coverage of commonly applied 16S rRNA gene amplicon
297 primer pairs for bacteria and archaea in ADs (Fig. 3). This information is highly
298 valuable when designing experiments, especially if targeting specific taxa.
299 Accordingly, we have determined the theoretical coverage for several commonly
300 applied primer pairs for all kingdom to species-level taxa in MiDAS 5 (Supplementary
301 Data 2). We found a fairly low coverage of the V1-V3 primer pair (perfect hits for
302 ≤79% of the bacterial FL-ASVs), which we commonly use due to its high phylogenetic
303 resolution [6,20]. We should therefore expect a significant bias when using this primer
304 pair. The V4 primer used here and in the Earth Microbiome project [37] showed good
305 coverage for both bacteria (perfect hits for 87% of the FL-ASVs) and archaea (perfect
306 hits for 98% of the FL-ASVs). However, a recently published primer pair for the V4
307 region, designed to improve coverage for Patescibacteria [38], showed even better
308 coverage for bacteria, achieving perfect hits for 97% of the FL-ASVs. Although this
309 primer pair does not target archaea, adding degeneracy at a single base in one of the
310 primers also provided coverage for archaea, with perfect hits for 98% of the FL-ASVs.
311 The exceptional coverage offered by this new primer pair leads us to recommend it for
312 the profiling of anaerobic digesters (ADs), despite its lower phylogenetic signal
313 compared to the V1-V3 primers. The V3-V5 primer pair, which was used here to target
314 archaea only, also had good coverage for archaea, though not as extensive as that of the
315 V4 primers, supporting the choice of the latter.
316
317 ***Effect of process and environmental factors on the AD microbiota***
318 Alpha diversity analyses showed that the rarefied (10,000 read per sample) ASV
319 richness and inverse Simpsons diversity in ADs were affected mainly by the primary
320 substrate type and the temperature in the ADs (Supplementary Fig. 5). Significantly
321 higher bacterial richness and diversity were observed for ADs treating surplus sludge
322 from WWTPs compared to the other types of substrates. This effect likely reflects the
323 extensive immigration of bacteria into the ADs with the surplus sludge [7,10,39]. A higher
324 richness and diversity were observed for bacteria in mesophilic ADs compared to
325 thermophilic ADs. A similar trend has previously been observed for full-scale ADs
326 treating manure [40,41], household waste [42], and surplus sludge from WWTPs [7].
327
328 Genus-level taxonomic beta-diversity was used to investigate the effect of process
329 conditions and geography on the overall microbiota in ADs using principal coordinate
330 analysis (PCoA) and permutational multivariate analysis of variance (PERMANOVA)
331 (Fig. 4). We used this approach because many of the important traits are categorical
332 (yes/no) and only conserved at lower taxonomic ranks (genus/species) [43]. Furthermore,
333 MiDAS 5 enabled us to classify almost all our ASVs at the genus-level, thereby
334 providing a comprehensive description of the microbiota. The PERMANOVA (Adonis
335 R^2 values) showed that the overall microbial community was mainly explained by the

primary substrate and to a lesser extent by temperature, continent, and digester technology (Fig. 4). This trend was observed for both bacteria and archaea. The percentage of total variation explained by each parameter was, except for the primary substrate, low, suggesting that the global AD microbiota represents a continuous distribution rather than distinct states, as also observed for the human gut microbiota [44] and WWTPs [20]. The pronounced effect of the primary substrates highlights that these substrates are distinct and abundant in microbes, particularly in the case of manure and wastewater sludge.

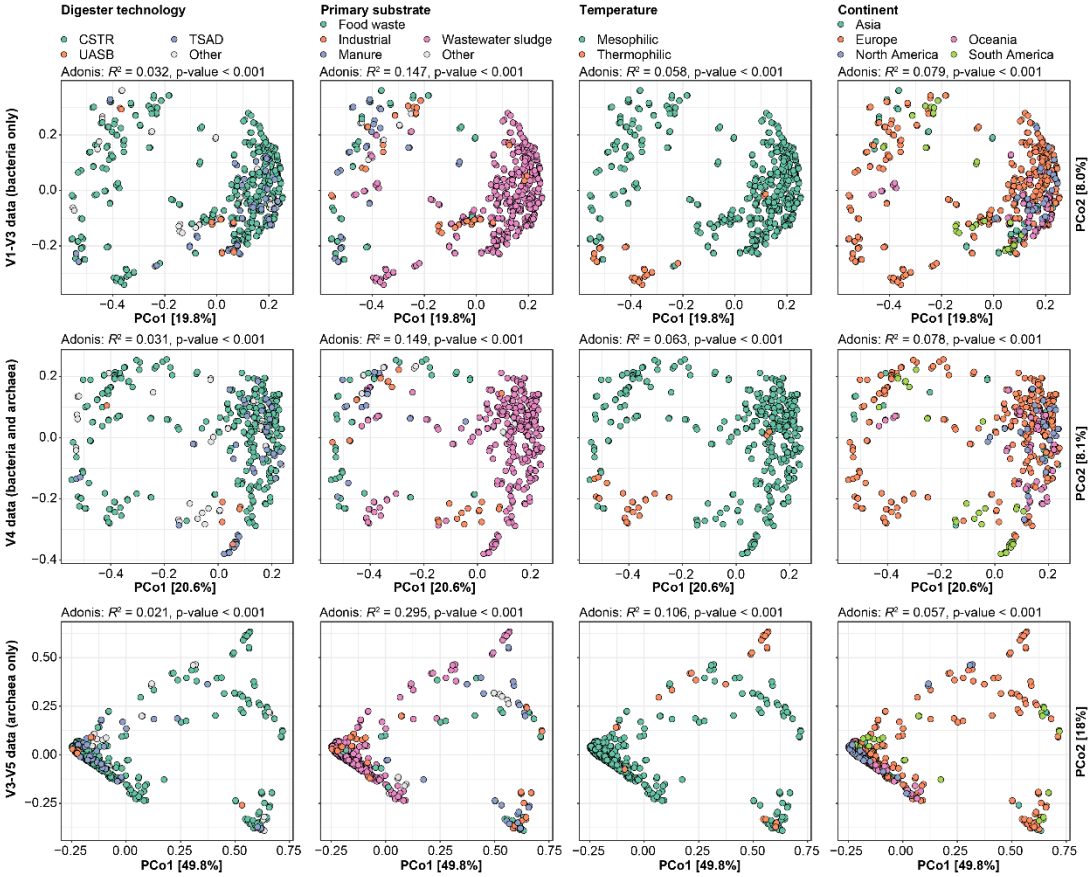

Fig. 4: Effects of process and environmental factors on the anaerobic digester microbiota. Principal coordinate analyses of Bray-Curtis beta-diversity for genera based on V1-V3 (bacteria), V4 (archaea and bacteria), and V3-V5 (archaea) amplicon data. Samples are colored based on metadata. The fraction of variation in the microbial community explained by each variable in isolation was determined by PERMANOVA (Adonis R^2-values). Exact p-values less than 0.001 could not be confidently determined due to the chosen number of permutations. CSTR: Continuous stirred-tank reactor; TSAD: Two-stage anaerobic digestion; UASB: Upflow anaerobic sludge blanket.

353 ***Core and conditional rare or abundant taxa in the global AD microbiota***
354 The global AD microbiota represents a huge microbial diversity. However, most
355 organisms only occur in very low abundance and are therefore unlikely to have any
356 quantitative impact on the overall metabolism and the process performance in ADs.
357 Analysis of core and conditionally rare or abundant taxa (CRAT) is a powerful
358 approach to identify the most important genera and species within a specific ecosystem
359 [20,28,45]. The CRAT may include taxa related to process disturbances, such as filamentous
360 microbes associated with foam formation, or taxa associated with the degradation of
361 special substrates found in, e.g., industrial waste.
362
363 We recently introduced and applied the following core and CRAT definitions in our
364 survey of the global microbiota of wastewater treatment plants: strict core (>0.1%
365 relative abundance in >80% of samples), general core (>0.1% relative abundance in
366 >50% of samples), loose core (>0.1% relative abundance in >20% of samples), and
367 CRAT (not part of the core, but present in at least one sample with a relative abundance
368 >1%) [20]. Here, we applied the same criteria to identify core and CRAT genera and
369 species in our global AD dataset. Because the primary substrate showed a strong effect
370 on the overall microbial community (Fig. 4), we determined the core and CRAT for
371 each individual substrate separately (Supplementary Data 3). Only mesophilic ADs
372 were examined for ADs treating food waste, industrial waste, and manure due to the
373 low number of thermophilic ADs sampled. Both mesophilic and thermophilic digesters
374 were examined for ADs treating wastewater sludge. To minimize the impact of primer
375 bias, we analyzed all three amplicon datasets and combined the results, including all
376 core and CRAT that were found in at least one of the datasets.
377
378 The core analysis revealed that most core genera were uniquely associated with specific
379 primary substrates and temperature range (Fig. 5a). However, there were also a
380 significant number of core genera shared across substrates (Fig. 5a). In contrast, very
381 few core species were shared between ADs treating different primary substrates (Fig.
382 5b). This fits well with similar results from a study of ADs in Belgium and Luxemburg
383 [13]. To define a 'most wanted' list for bacteria and archaea in ADs globally, we linked
384 each core and CRAT to their highest-ranking category across primary substrates,
385 process temperatures, and primer pair (Supplementary Data 3). The resulting list
386 contained 501 core genera (75 strict, 117 general, and 309 loose) and 191 CRAT genera.
387 The strict core genera included 11 known methanogens and four known syntrophs (*Ca*.
388 Phosphitivorax, *Smithella*, *Syntrophomonas*, *Syntrophorhabdus*). At the species-level,
389 we identified 565 core (29 strict, 126 general, and 410 loose) and 448 CRAT species.
390 The strict core species included two methanogens (*Methanobrevibacter smithii* and
391 *Methanothermobacter* midas_s_3958) and one syntroph (*Syntrophomonas*
392 midas_s_90707). It is worth noting that a large fraction of the taxa observed in
393 ADs does not grow in the digesters, but only occurs because the taxa are in high
394 abundance in the feed [7,10,39]. Previous published data from Danish ADs treating

wastewater sludge [7] classified 45 (9.0%) of the core genera observed in this study as non-growing (<20% of ASVs belonging to the specific taxa were classified as growing), whereas 393 (78.4%) were classified as growing. A similar analysis of core species classified 45 (8.0%) as non-growing and 391 (69.2%) as growing. However, it remains to be determined if these numbers also translate to global ADs.

Many core and CRAT represent MiDAS placeholder taxa

A large proportion of core and CRAT identified was classified as MiDAS *de novo* taxa. At the genus-level, 272/501 (54%) of the core genera and 119/191 (62%) of the CRAT genera had only MiDAS placeholder names, and at the species-level, the proportion was even higher. Here placeholder names were assigned to 514/565 (91%) of the core species and 422/448 (94%) CRAT species. These proportions are similar to those observed for the global microbiota in activated sludge [20] and reveals the importance of a taxonomic framework that can handle uncultured taxa which have not yet been officially classified.

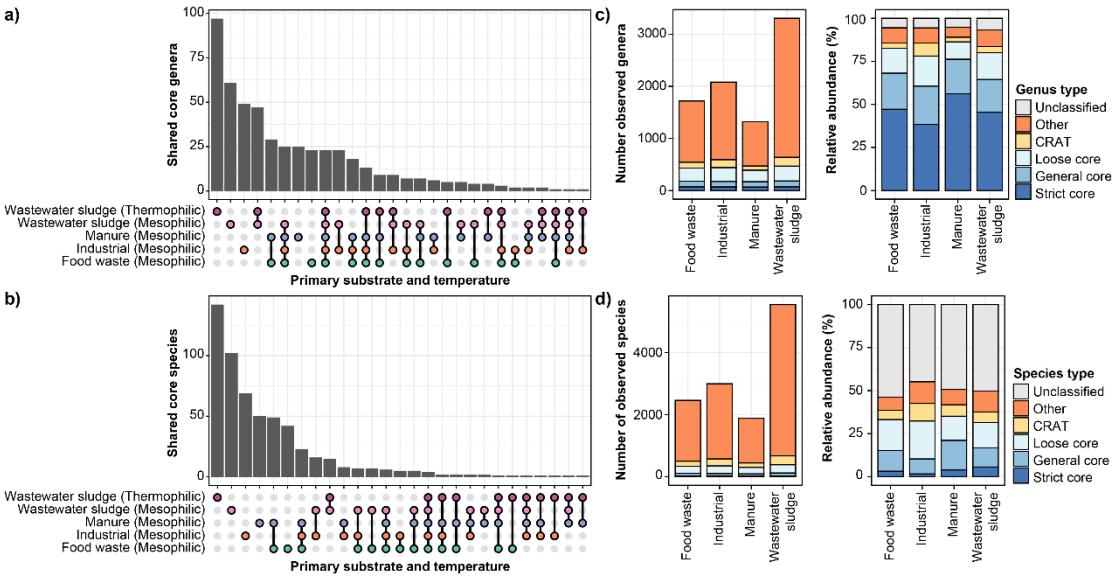

Fig. 5: Core and conditionally rare or abundant taxa (CRAT) in anaerobic digesters globally. a) and b) UpSet plots displaying the number of shared core genera and species, respectively, across ADs treating different primary substrates and operating at different temperatures. c) and d) Number of observed genera and species, respectively, and their abundance in mesophilic ADs treating different primary substrates based on V4 amplicon data (bacteria and archaea). Values for genera and species are divided into strict core, general core, loose core, CRAT, other taxa, and unclassified ASVs based on the most wanted list (Supplementary Data 3). The relative abundance of different groups was calculated based on the mean relative abundance of individual genera or species across samples. Similar figures for V1-V3 (bacteria only) and V3-V5 (archaea only) amplicons data can be found in Supplementary Fig. 6.

The global AD microbiota is dominated by core and CRAT taxa

Despite only accounting for a minor fraction of the total diversity in the ADs examined, the core and CRAT represented most of the microbes ~~in the ecosystem~~ according to relative amplicon read abundance (Fig. 5c, 5d, Supplementary Fig. 6). The core and CRAT genera accounted for 85-92% (V1-V3), 84-89% (V4), and 96-99% (V3-V5) of the accumulated read abundance in mesophilic ADs depending on primary substrates. The remaining fractions consisted mainly of ASVs unclassified at the genus level, and genera present in very low abundance, presumably with minor importance for the AD performance.

For the species level, the core and CRAT represented 53-61% (V1-V3), 38-43% (V4), and 18-47% (V3-V5) accumulated read abundance depending on the primary substrate. The remaining fractions were mainly composed of ASVs, which could not be classified at the species level, probably due to insufficient phylogenetic resolution in the short-read amplicons [6,31]. The lack of species-level classification was especially pronounced for the archaeal V3-V5 ASVs in ADs treating industrial waste, manure, and wastewater sludge (Supplementary Fig. 6).

The large relative abundance of core and CRAT in the global AD microbiota suggests that we can explain a substantial part of the metabolic processes in ADs, if we understand the physiology and metabolic potential of these taxa.

Global diversity of archaea reveals new potential methanogens

As methanogenic archaea are ultimately responsible for the generation of methane in ADs, we examined the global diversity of archaea in all samples based on V4 (Fig. 6) and V3-V5 amplicon data (Supplementary Fig. 7). The V4 amplicon data, encompassing both archaea and bacteria, showed that the archaeal reads constituted 5.6%±4.4% for ADs treating food waste, 6.8%±4.4% for manure, 6.4%±2.5% for wastewater sludge, and 13.7%±11.1% for industrial waste. Many of the abundant archaea represented well-known methanogens. However, we also observed several abundant genera, only classified based on the MiDAS placeholder taxonomy, affiliated to orders and families of known methanogens. These include midas_g_91627 and midas_g_8154, which represent new families within the orders Methanomicrobiales and Methanofastidiosales, respectively, and midas_g_90473 and midas_g_93310, representing new genera within Methanomassiliicoccaceae and Methanospirillaceae, respectively. In addition, we observed two abundant MiDAS placeholder genera (midas_g_90791 and midas_g_97217) that represent a new order within the class *Ca.* Bathyarchaeia. Members of this class can have a versatile metabolism, and some encode the key methanogenic enzyme methyl-coenzyme M reductase (MCR) [46,47]. Targeted metagenomics and assembly of metagenome-assembled genomes (MAGs) should be applied to confirm the methanogenic potential of these new potential

methanogens, and our amplicon datasets provide insight into where these taxa occur in high abundance.

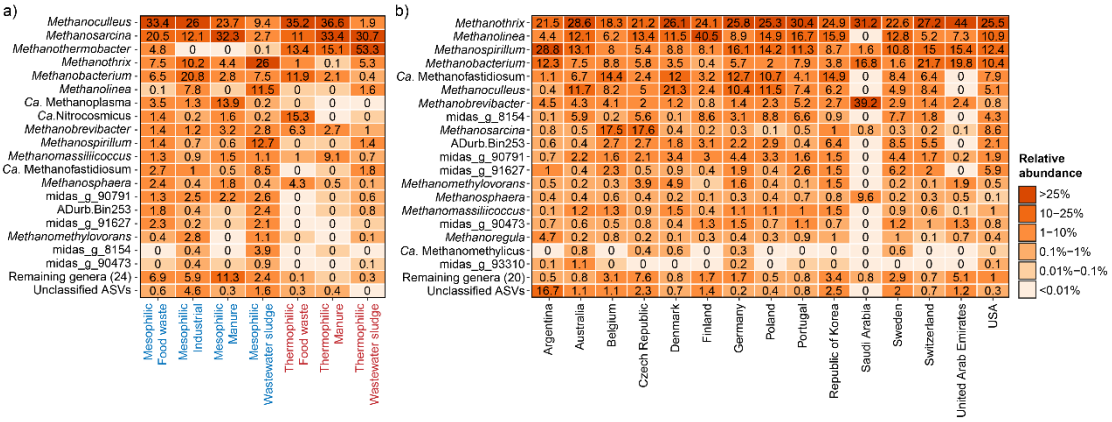

Fig. 6: Top 25 archaeal genera based on V4 amplicon data. The percent relative abundance represents the mean abundance relative to all archaea across a) different temperature range and primary substrates, and b) different countries considering only mesophilic ADs treating mainly wastewater sludge.

The methanogenic community composition was clearly affected by the primary substrate and whether the digestion was performed at mesophilic or thermophilic conditions (Fig. 6a, Supplementary Fig. 7a). The most common methanogens across substrates and temperatures were *Methanoculleus*, *Methanosarcina*, *Methanothermobacter*, and *Methanothrix*. *Methanothermobacter* was as expected most abundant in thermophilic ADs. However, to our surprise, it also occurred in high relative abundance in several mesophilic reactors treating mainly food waste. We were not able to explain their occurrences in these ADs based on the available metadata for the plants, but future studies might shed light on the underlying mechanisms or environmental factors that enable this unexpected distribution.

Because most of our samples originated from mesophilic reactors treating wastewater sludge, we examined the diversity of methanogens across countries in these ADs (Fig. 6b, Supplementary Fig. 7b). This analysis revealed that the same genera were dominating across the world. The most common methanogens in these ADs were *Methanothrix*, *Methanolinea*, *Methanospirillum*, *Methanobacterium*, and the recently discovered *Ca*. Methanofastidiosum [48]. Next, we examined if the methanogens were also conserved at higher phylogenetic resolution. As many archaeal ASVs could not be classified at the species-level, we examined the global diversity at the ASV-level (Supplementary Fig. 8). We found that the vast majority of the abundant ASVs occurred globally. The significant similarity of methanogens across various regions indicates substantial potential for global knowledge transfer concerning their management and utilization.

Among the highly abundant archaea, we also observed an ammonia oxidizing archaeon (AOA) from the genus *Ca*. Nitrosocosmicus [49], which was especially abundant in thermophilic ADs treating food waste. This is surprising and may indicate that they also

have an anaerobic physiology which should be investigated further. Another abundant archaeon was the *Ca.* Diapherotrites ADurb.bin253 belonging to the order Woesearchaeales which are characterized by ultra-small genomes and an anaerobic and parasitic/fermentation-based lifestyle [50].

Global diversity of syntrophic bacteria

Syntrophic bacteria play a vital role in ADs by converting substrates, such as short-chain fatty acids, into acetate, H_2, and formate [29,51,52]. These compounds serve as substrates or reducing equivalents for methanogens, which in turn produce methane and CO_2. This obligately mutualistic metabolism is crucial because the syntrophs can only oxidize substrates and sustain growth under anaerobic conditions if the methanogens rapidly consume the products to maintain them at very low concentrations [51,53]. Due to the fastidious metabolism, syntrophs are usually present in low abundance, and can easily become the bottleneck in the anaerobic digestion process [7,8]. Accordingly, we investigated the global diversity of this functional guild in the ADs sampled (Fig. 7, Supplementary Fig. 9).

A clear effect of the primary substrates and digester temperature was observed on the composition and abundance of syntrophic genera in the digesters (Fig. 7a, Supplementary Fig. 9a). The most abundant genus across substrates and temperature was *Syntrophaceticus*, despite being barely detected in ADs treating wastewater sludge. The type strain of this genus, *S. schinkii* Sp3^T, is an acetate-oxidizing syntroph that thrives, and has a competitive advantage, under high ammonium concentrations (up to 8,400 mgN/L) [54,55]. The lack of *Syntrophaceticus* in ADs treating wastewater sludge may therefore be explained by lower ammonium concentrations in these ADs (1,617±4,312 mgN/L, n=145) compared to those treating food waste (2,913±1,681 mgN/L, n=33), and manure (3,449±933 mgN/L, n=18).

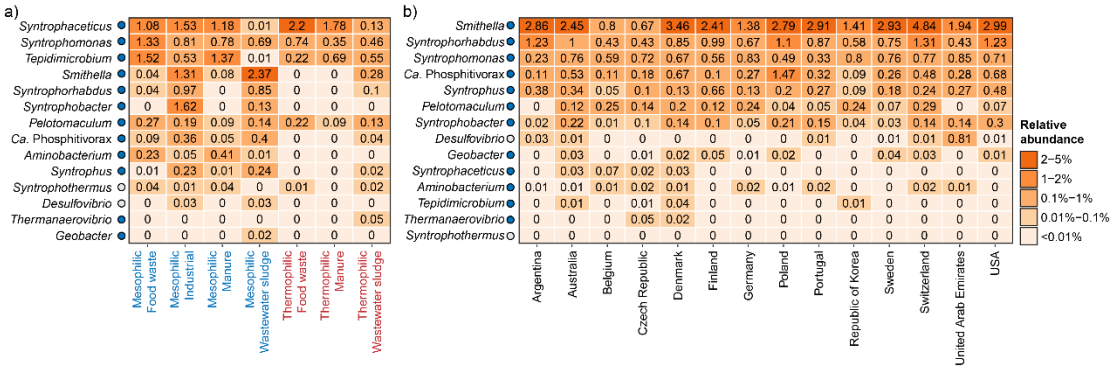

Fig. 7: Global diversity of syntrophs based on V1-V3 amplicon data. The percent relative abundance represents the mean for genera across a) different temperature range and primary substrates, and b) different countries considering only mesophilic ADs treating mainly wastewater sludge. Colored circles next to the genus labels indicate whether the genera have previously been identified as growing in ADs at Danish WWTPs according to Jiang *et al.* [7]. Blue: >50% of ASVs classified as growing; Yellow: 20-50% of ASVs classified as growing. Red: <20% of ASVs classified as growing. Gray: No information available for the specific genus.

532 *Syntrophomonas*, the second most abundant genus, was common in all AD types
533 investigated, indicating a broader ecological niche. Isolated representatives from this
534 genus can grow syntrophically via β-oxidation of saturated fatty acids of various lengths
535 (C4-C18, depending on strain) [56–59], and they are therefore likely important for the
536 conversion of long-chain fatty acids in ADs. Among the abundant syntrophs,
537 *Tepidimicrobium,* a member of the order Clostridiales, was also observed in all AD
538 types except mesophilic ADs treating wastewater sludge. The exact metabolism of
539 *Tepidimicrobium* in ADs remains to be determined, however all isolated representatives
540 can degrade proteinaceous compounds and some species can also use carbohydrates [60].
541 Furthermore, *Tepidimicrobium* has been proposed to grow syntrophically by direct
542 interspecies electron transfer (DIET) with *Methanothermobacter* in a process like that
543 observed for *Geobacter* [61]. Accordingly, it is likely that the *Tepidimicrobium* acts as a
544 syntrophic primary degrader in the ADs targeting mainly proteins, carbohydrates, and
545 derivatives.
546
547 Finally, we observed a high abundance of the genus *Smithella* in mesophilic ADs
548 treating industrial waste, manure, and wastewater sludge. The type strain *S. propionica*
549 LYP^T is a propionate oxidizing syntroph, which uses a unique dismutation pathway in
550 which propionate is first converted to acetate and butyrate, and the latter is hereafter β-
551 oxidized syntrophically to acetate and hydrogen [62,63]. Calculations of Gibbs free energy
552 for this special propionate metabolism indicates a higher tolerance toward elevated
553 hydrogen concentrations [64], which could explain why some *Smithella* prevail in certain
554 ADs. However, *Smithella* has also been implicated in the syntrophic degradation of
555 long-chain alkanes [65,66], which could reflect a more versatile metabolism.
556
557 When investigating geographical diversity of syntrophic fatty acid oxidizing bacteria
558 in mesophilic ADs treating wastewater sludge, a similar pattern was observed across
559 countries (Fig. 7b, Supplementary Fig. 9b). *Smithella*, was generally the dominating
560 syntroph. However, *Syntrophomonas*, *Syntrophorhabdus*, *Ca*. Phosphitivorax, and
561 *Syntrophus* also occurred at a high relative abundance in almost all countries. Isolates
562 of *Syntrophorhabdus,* including the type strain *S. aromaticus* UI^T, are syntrophic
563 fermenters of aromatic compounds and may accordingly play an important role in the
564 detoxification of these substrates in ADs [67,68]. *Ca*. Phosphitivorax was recently
565 discovered as a butyrate degrading syntroph by genome-resolved meta-transcriptomics
566 in a digester treating wastewater sludge [52], and *Syntrophus* participates in the
567 degradation of fatty acids and aromatics [69,70]. Overall, the results suggest a complex
568 syntrophic degradation process, which involves multiple genera with different substrate
569 specificities.
570
571 To gain additional insight into the global diversity of syntrophs, we also investigated
572 the species-level diversity across mesophilic digesters treating wastewater sludge
573 (Supplementary Fig. 10). We observed a large species diversity among most of the

abundant syntrophic genera. Furthermore, we found that the most abundant species in the ADs were often distinct from the isolated representatives, which prompts for further investigations into the metabolic potential of syntrophs *in situ*.

Global diversity of filamentous bacteria

Foaming is a common operational problem in ADs and has a strong negative impact on process performance resulting in considerable costs. Both abiotic and biotic factors are involved in foaming [71]. The abiotic factors include high loading rates of surfactants (oil, grease, fatty acids, detergent, proteins, and particulate matter) and biosurfactants produced by microbes in the digester [72]. The biotic factors cover increased abundance of hydrophobic, filamentous microorganisms that can interact with, and stabilize, gas bubbles in the foam [71,73]. To gain further insight into potential foam forming microbes, we examined the global diversity of known filamentous bacteria in ADs (Fig. 8, Supplementary Fig. 11).

The diversity and mean relative abundance of known filamentous organisms were generally low in the ADs examined except for those treating wastewater sludge (Fig. 8a, Supplementary Fig. 11). However, the increased diversity and abundance in the latter are to a large extent the result of passive immigration from the fed surplus sludge. However, most of these are unable to grow in the ADs [7]. *Anaerolinea*, *Ca*. Brevefilum, and *Trichococcus* were common across ADs treating all primary substrates (Fig. 8a, Supplementary Fig. 11), whereas *Ca*. Microthrix and *Ca.* Promineofilum were mainly observed in ADs treating wastewater sludge. Many of the Chloroflexi genera found here were also observed in a recent meta-analysis of amplicon data from 17 studies representing 62 ADs [74]. Several of the abundant filamentous genera, including *Ca*. Microthrix and *Ca*. Brevefilum, were previously found to correlate with the foaming potential of full-scale digester sludge from mesophilic ADs at WWTPs [73]. *Ca*. Brevefilum seems especially interesting as it grows well in ADs [7,75].

The species-level diversity was generally low for the filamentous bacteria (Supplementary Fig. 12). *Ca*. Brevefilum was dominated by *Ca*. B. fermentans, *Trichococcus* by midas_s_4, *Ca*. Microthrix by *Ca.* M. parvicella and *Ca.* M. subdominans, and *Gordonia* by *G. defluvii* and *G. amarae*. *Ca.* Promineofilum was dominated by *Ca.* P. glycogenico, but a few MiDAS placeholder species, were also commonly observed. The low species-level diversity of potential foam-forming bacteria suggests that it may be feasible to develop and implement universal mitigation strategies for these bacteria in ADs worldwide.

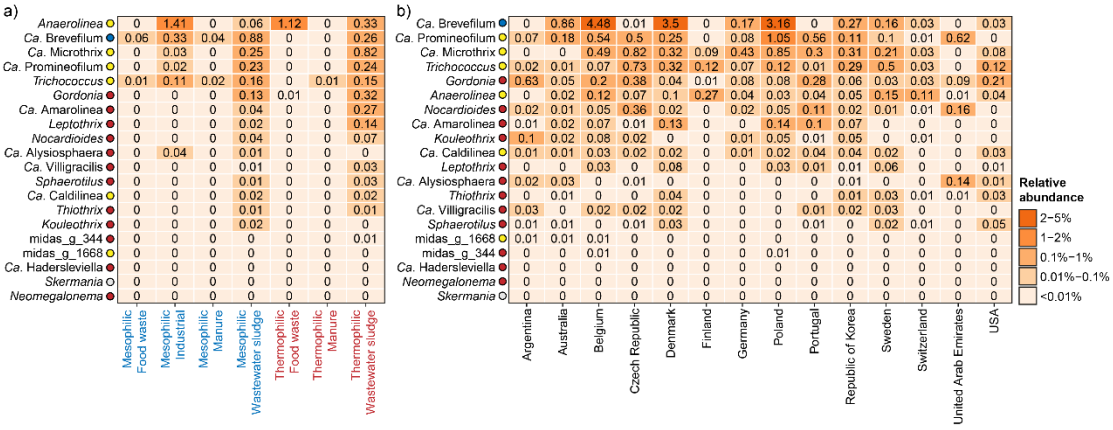

Fig. 8: Global diversity of known filamentous organisms based on V1-V3 amplicon data. The percent relative abundance represents the mean for genera across a) different temperature range and primary substrates, and b) different countries considering only mesophilic ADs treating mainly wastewater sludge. Colored circles next to the genus labels indicate whether the genera have previously been identified as growing in ADs at Danish WWTPs according to Jiang *et al.* [7]. Blue: >50% of ASVs classified as growing; Yellow: 20-50% of ASVs classified as growing. Red: <20% of ASVs classified as growing. Gray: No information available for the specific genus.

Conclusion and perspectives

MiDAS 5 was made possible thanks to a huge collaborative effort from experts worldwide, who contributed to the project by sampling and providing metadata for ADs in their respective countries. Building on the success of its predecessor, MiDAS 4, this latest expansion covers ASV-resolved, full-length 16S rRNA gene references from numerous ADs from all parts of the globe covering different operations parameters and different substrates. This expanded database provides greatly improved coverage for AD-specific taxa and a strongly needed taxonomy for uncultured lineages, which lacks official taxonomic classification. As such, it will be an invaluable resource for researchers and AD professionals, providing them with a common point of reference to facilitate knowledge sharing and pave the way for a comprehensive understanding of the AD microbiome.

Our in silico 16S rRNA gene primer evaluation based on the MiDAS 5 database revealed that the coverage of commonly applied primer pairs varies significantly, with some having low coverage and potential bias towards certain taxa. Because the primer coverage was evaluated for all taxa in the MiDAS 5 database at all taxonomic ranks, it provides a solid foundation for designing experiments and targeting specific taxa in future studies. For general microbial profiling of ADs, we would recommend the use of the newly improved universal V4 primer pair [38], as it show excellent coverage for both archaea and bacteria in the AD ecosystem.

641 Although the total microbial diversity in ADs is huge, importantly, we showed that less
642 than 1000 genera and species accounted for most of the microbes in the AD ecosystem.
643 By focusing on the fraction of these abundant and common microbes that can grow in
644 the AD systems, we will be able to explain most of the microbial processes that occur
645 in the anaerobic digestion process. This list of "Most Wanted" organisms contain
646 species that should be prime targets for future in situ studies and the reconstruction of
647 MAGs. These genomes can then be annotated to provide additional details about their
648 potential metabolic pathways and roles in the AD ecosystem [15,16,76–78].
649
650 The global survey of the ADs microbiota using three different primer pairs provided a
651 unique insight into the global diversity of individual AD taxa and clues into the
652 environmental and operational factors that define their ecological niches. This
653 information will be invaluable in the development of future microbiome management
654 strategies and improved sustainability of the field of anaerobic digestion.
655
656 To enhance knowledge dissemination, we have updated the MiDAS Field Guide
657 available at www.midasfieldguide.org. This dynamic resource allows users to delve
658 into specifics related to the physiology, morphology, and ecology of genera listed in
659 the MiDAS database. Additionally, it offers country-specific data on the prevalence of
660 all MiDAS genera and species in WWTPs and ADs. Finally, it provides information on
661 the availability of fluorescence in situ hybridization probes and reference genomes,
662 paving the way for subsequent research endeavors.

663 **Methods:**

664 ***Sampling and metadata collection***

665 To facilitate sampling of ADs worldwide, we established the MiDAS Global
666 Consortium for Anaerobic Digesters, which consists of 25 anaerobic digestion experts
667 in 19 countries. Members of the consortium acted as national sampling coordinators
668 and were in direct contact with the ADs. Two samples were obtained from each AD
669 and shipped on ice to the sampling coordinators. For each replicate, 2 mL sample was
670 preserved in 2 mL RNAlater (Invitrogen), stored at 4°C until all national samples were
671 collected (usually within a few days), and then shipped to Aalborg University with
672 cooling elements. Upon arrival, the samples were separated into aliquots that were
673 prepared for nucleic acid purification. Metadata associated with each AD was also
674 obtained by the sampling coordinators and is provided as Supplementary Data 1.
675 Minimum information from all ADs included continent, country, GPS coordinates,
676 sampling date, temperature in the digester ("Mesophilic" (≤45°C) or "Thermophilic"
677 (50-60°C)), primary substrate ("Wastewater sludge", "Industrial", "Food waste",
678 "Manure", or "Other"), and digester technology ("Two-stage digester (TSAD)",
679 "Continuous Stirred Tank Reactors (CSTR)", "Upflow anaerobic sludge blanket
680 (UASB)", or "Other").

681

682 ***General molecular methods***

683 All commercial kits were used according to the protocols provided by the manufacturer
684 unless otherwise stated. The concentration and quality of nucleic acids were determined
685 using a Qubit 3.0 fluorometer (Thermo Fisher Scientific) and an Agilent 2200
686 Tapestation (Agilent Technologies), respectively.

687

688 ***Nucleic acid purification***

689 DNA was purified using a custom plate-based extraction protocol based on the
690 FastDNA spin kit for soil (MP Biomedicals). The protocol is available at
691 www.midasfieldguide.org (aau_ad_dna_v 2.0). RNAlater preserved samples were
692 thawed and homogenized using a Heidolph RZR 2020 laboratory stirrer. 20 µL of
693 sample was resuspended in 300 µL PBS and transferred to Lysing Matrix E barcoded
694 tubes (MP Biomedicals). 40 µL of MT buffer was added and lysis was performed by
695 bead beating in a FastPrep-96 bead beater (MP Biomedicals) (3x 120 s, 1800 rpm with
696 2 min incubation on ice between cycles). The samples were centrifuged (3,486 ×g, 10
697 min) and 200 µL supernatant was transferred to a 96-well PCR-plate. 50 µL Protein
698 Precipitation Solution (PPS) was mixed with each sample, which was then centrifuged
699 again. 150 µL supernatant was cleaned-up using 100 µL CleanNGS beads with elution
700 into 60 µL of nuclease-free water. 40 µL of the purified DNA was transferred to a new
701 96-well plate and stored at -80°C.

702 ***Full-length 16S rRNA gene library preparation, sequencing, and processing***
703 Full-length 16S rRNA gene sequencing was carried out using high-accuracy, long-read
704 amplicon sequencing using unique molecular identifiers (UMIs) and PacBio circular
705 consensus sequencing (CCS) [79]. Oligonucleotides used can be found in Supplementary
706 Table 1. Bacterial and archaeal 16S rRNA genes were UMI-tagged using overhang
707 primers based on the 27F and 1391R [80] and SSU1ArF and SSU1000ArR [34] primer pairs,
708 respectively. These primers have shown excellent coverage for the known bacterial and
709 archaeal diversity in silico [34,80].
710
711 *Addition of UMI-tags by overhang PCR:* Adaptors containing UMIs, and defined
712 primer binding sites were added to each end of the bacterial and archaeal 16S rRNA
713 genes by PCR. The reaction contained 20 µL of 5x SuperFi Buffer (Invitrogen), 2 µL
714 of 10 mM dNTP mix, 5 µL of 10 µM f16S_pcr1_fw, 5 µL of 10 µM f16S_pcr1_rv, 1
715 µL of 2 U/µL Platinum SuperFi DNA polymerase (Invitrogen), 100 ng of pooled
716 template DNA (from all ADs), and nuclease-free water to 100 µL. The reaction was
717 incubated with an initial denaturation at 98°C for 30 s followed by 2 cycles of
718 denaturation at 98°C for 20 s, annealing at 55°C for 30 s, and extension at 72°C for 45
719 s, and then a final extension at 72°C for 5 min. The sample was purified using 0.6x
720 CleanNGS beads and eluted in 20 µL nuclease-free water.

721 *Primary library amplification:* The tagged 16S rRNA gene amplicons were amplified
722 using PCR to obtain enough product for quantification. The reaction contained 19 µL
723 of UMI-tagged sample, 20 µL 5x SuperFi buffer (Invitrogen), 2 µL of 10 mM dNTP, 5
724 µL of 10 µM f16S_pcr2_fw, 5 µL of 10 µM f16S_pcr2_rv, 48 µL nuclease-free water,
725 and 1 µL 2U/µL Platinum SuperFi DNA polymerase (Invitrogen). The reaction was
726 incubated with an initial denaturation at 98°C for 30 s followed by 15 cycles of
727 denaturation at 98°C for 20 s, annealing at 60°C for 30 s, and extension at 72°C for 45
728 s and then a final extension at 72°C for 5 min. The PCR product was purified using
729 0.6x CleanNGS beads and eluted in 11 µL nuclease-free water. The amplicons were
730 validated on a Genomic screentape and quantified with the Qubit dsDNA HS assay kit.

731 *Clonal library amplification:* Tagged amplicon libraries were diluted to approximately
732 250,000 molecules/µL and amplified by PCR to obtain clonal copies of each uniquely
733 tagged amplicon molecule. Three libraries were made for the bacterial 16S rRNA genes
734 and one for archaea. The PCR reactions contained 1 µL diluted primary library, 20 µL
735 5x SuperFi buffer (Invitrogen), 2 µL of 10 mM dNTP, 5 µL of 10 µM f16S_pcr2_fw,
736 5 µL of 10 µM f16S_pcr2_rv, 66 µL nuclease-free water, and 1 µL 2U/µL Platinum
737 SuperFi DNA polymerase (Invitrogen). The reaction was incubated with an initial
738 denaturation at 98°C for 30 s followed by 25 cycles of denaturation at 98°C for 20 s,
739 annealing at 60°C for 30 s, and extension at 72°C for 45 s and then a final extension at
740 72°C for 5 min. The PCR product was purified using 0.6x CleanNGS beads and eluted
741 in 20 µL nuclease-free water. The amplicons were validated on a Genomic screentape
742 and quantified with the Qubit dsDNA HS assay kit.

743 *PacBio CCS sequencing:* The four clonal libraries were sent to Admera Health
744 (Plainfield, NJ, USA) for PacBio library preparation and sequencing. Here amplicons
745 were incubated with T4 polynucleotide kinase (New England Biolabs) following the
746 manufacturer's instructions, and sequencing library prepared using SMRTbell Express
747 Template Preparation kit 1.0 following the standard protocol. Sequencing was
748 performed using 4x SMRT cells on a Sequel II using a Sequel II Sequencing kit 1.0,
749 Sequel II Binding and Int Ctrl kit 1.0 and Sequel II SMRT Cell 8M, following the
750 standard protocol with 1 h pre-extension and 15 h collection time (Pacific Biosciences).

751 *Bioinformatic processing:* CCS reads were generated from raw PacBio data using CCS
752 v.3.4.1 (https://github.com/PacificBiosciences/ccs) with default settings. UMI
753 consensus sequences (consensus_raconx3.fa) were obtained using the longread_umi
754 script (https://github.com/SorenKarst/longread_umi) [79] using the following options:
755 pacbio_pipeline, -v 3, -m 1000, -M 2000, -s 60, -e 60, -f
756 CAAGCAGAAGACGGCATACGAGAT, -F AGRGTTYGATYMTGGCTCAG
757 (bacteria) or TCCGGTTGATCCYGCBRG (archaea), -r
758 AATGATACGGCGACCACCGAGATC, -R GACGGGCGGTGWGTRCA (bacteria)
759 or GGCCATGCAMYWCCTCTC (archaea), and -c 3. The UMI-consensus reads were
760 oriented based on the SILVA 138.1 SSURef NR99 database using the usearch
761 v.11.0.667 -orient command and trimmed between the 27f and 1391r (bacteria) or
762 SSU1ArF and SSU1000ArR (archaea) primer binding sites using the trimming function
763 in CLC genomics workbench v. 20.0. Sequences without both primer binding sites were
764 discarded. The trimmed high-fidelity reads were processed with AutoTax v. 1.7.4 [6] to
765 create full-length 16S rRNA gene amplicon sequence variants (FL-ASV) and these
766 were added to the MiDAS 4.8.1 reference database [20] to create MiDAS 5.0. Subsequent
767 updates to MiDAS 5.2 were made to accommodate taxonomic updates (see the release
768 change logs for details).

769 ***Short-read amplicon sequencing***
770 V1-V3 amplicons were made using the 27F (5'-AGAGTTTGATCCTGGCTCAG-3')
771 [81] and 534R (5'-ATTACCGCGGCTGCTGG-3') [82] primers with barcodes and Illumina
772 adaptors (IDT) [83]. 25 µL PCR reactions in duplicate were run for each sample using 1X
773 PCRBIO Ultra Mix (PCR Biosystems), 400 nM of both forward and reverse primer,
774 and 10 ng template DNA. PCR conditions were 95°C, for 2 min followed by 20 cycles
775 of 95°C for 20 s, 56°C for 30 s, and 72°C for 60 s, followed by a final elongation at
776 72°C for 5 min. PCR products were purified using 0.8x CleanNGS beads and eluted in
777 25 µL nuclease-free water.
778
779 V3-V5 amplicons were made using the Arch-340F (5'-CCCTAHGGGGYGCASCA-
780 3') and Arch-915R (5'-GWGCYCCCCCGYCAATTC-3') primers [84]. 25 µL PCR
781 reactions in duplicate were run for each sample using 1X PCRBIO Ultra Mix (PCR
782 Biosystems), 400 nM of both forward and reverse primer, and 10 ng template DNA.
783 PCR conditions were 95°C, for 2 min followed by 30 cycles of 95°C for 15 s, 55°C for

784 15 s, and 72°C for 50 s, followed by a final elongation at 72°C for 5 min. PCR products
785 were purified using 0.8x CleanNGS beads and eluted in 25 µL nuclease-free water. 2
786 µL of purified PCR product from above was used as template for a 25 µL Illumina
787 barcoding PCR reaction containing 1x PCRBIO Reaction buffer, 1 U PCRBIO HiFi
788 Polymerase (PCR Biosystems) and 10 µL of Nextera adaptor mix (Illumina). PCR
789 conditions were 95°C, for 2 min, 8 cycles of 95°C for 20 s, 55°C for 30 s, and 72°C for
790 60 s, followed by a final elongation at 72°C for 5 min. PCR products were purified
791 using 0.8x CleanNGS beads and eluted in 25 µL nuclease-free water.
792
793 V4 amplicons were made using the 515F (5'-GTGYCAGCMGCCGCGGTAA-3') [82]
794 and 806R (5'-GGACTACNVGGGTWTCTAAT-3') [85] primers. 25 µL PCR reactions
795 in duplicate were run for each sample using 1X PCRBIO Ultra Mix (PCR Biosystems),
796 400 nM of both forward and reverse primer, and 10 ng template DNA. PCR conditions
797 were 95°C, for 2 min followed by 30 cycles of 95°C for 15 s, 55°C for 15 s, and 72°C
798 for 50 s, followed by a final elongation at 72°C for 5 min. PCR products were purified
799 using 0.8x CleanNGS beads and eluted in 25 µL nuclease-free water. 2 µL of purified
800 PCR product from above was used as template for a 25 µL Illumina barcoding PCR
801 reaction as described for the V3-V5 amplicons.
802
803 16S rRNA gene V1-V3, V3-V5, and V4 amplicon libraries were pooled separately in
804 equimolar concentrations and diluted to 4 nM. The amplicon libraries were paired-end
805 sequenced (2 × 300 bp) on the Illumina MiSeq using v3 chemistry (Illumina, USA). 10
806 to 20% PhiX control library was added to mitigate low diversity library effects.
807
808 ***Processing of short-read amplicon data***
809 Usearch v.11.0.667 [86] was used for processing of 16S rRNA gene amplicon data and
810 for read mapping. V1-V3 forward and reverse reads were merged using the usearch -
811 fastq_mergepairs command, filtered to remove phiX sequences using usearch -
812 filter_phix, and quality filtered using usearch -fastq_filter with -fastq_maxee 1.0.
813 Dereplication was performed using -fastx_uniques with -sizeout, and amplicon
814 sequence variants (ASVs) were resolved using the usearch -unoise3 command [87]. An
815 ASV-table was created by mapping the quality filtered reads to the ASVs using the
816 usearch -otutab command with the -zotus and -strand plus options. Taxonomy was
817 assigned to ASVs using the usearch -sintax command with -strand both and -
818 sintax_cutoff 0.8 options. Mapping of ASVs to reference databases was done with the
819 usearch -usearch_global command and the -id 0, -maxaccepts 0, -maxrejects 0, -
820 top_hit_only, and -strand plus options.
821
822 16S rRNA gene V3-V5 forward reads (reverse reads in relation the 16S rRNA gene)
823 were filtered to remove phiX sequences using usearch -filter_phix, trimmed to remove
824 primers and obtain a fixed length of 250 bp using -fastx_truncate with -stripleft -17 and
825 trunclen 250, reverse complemented with usearch -fastx_revcomp, and quality filtered

using usearch -fastq_filter with -fastq_maxee 1.0. Subsequent processing was like that for the V1-V3 amplicons.

16S rRNA gene V4 forward reads (reverse reads in relation the 16S rRNA gene) were trimmed with cutadapt v.2.8 [88] based on the V4 primers with the -g ^GGACTACHVGGGTWTCTAAT...TTACCGCGGCKGCTGGCAC and --discard-untrimmed options. The trimmed reads, which span the entire V4 amplicon, were reverse complemented with usearch -fastx_revcomp, and quality filtered using usearch -fastq_filter with -fastq_maxee 1.0. Subsequent processing was like that for the V1-V3 amplicons.

In silico *primer evaluation*

The specificity of commonly used amplicon primers was determined for each FL-ASV using the analyze_primers.py script from Primer Prospector v. 1.0.1 [89]. The specificity of primer sets was defined based on the overall weighted scores (OWS) for the primer with the highest score as follows: Perfect hit (OWS = 0), partial hit (OWS > 0, and ≤ 1), poor hit (OWS > 1). The percentage of perfect hits were calculated in R for all taxa in MiDAS 5.

Microbial community analyses

Short-read amplicon data was analyzed with R v.4.3.2 [90] through RStudio IDE v.2023.12.1 [91], with the tidyverse v.2.0.0 (https://www.tidyverse.org/), vegan v.2.6-4 [92], maps v.3.4.2 [93], data.table v.1.14.10 [94], FSA v.0.9.5 [95], rcompanion v. 2.4.35 [96], patchwork v.1.1.3 [97], ggupset v.0.3.0 [98] and Ampvis2 v.2.8.6 [99] packages.

The microbial community analyses were performed based on all three 16S rRNA gene short-read amplicon dataset (V1-V3, V3-V5, and V4). Samples with less than 10,000 reads and those lacking information about digester technology, primary substrate, and temperature in the digester were discarded from the analyses. After filtration, 547 V1-V3, 542 V3-V5, and 430 V4 samples remained.

Associations between the AD microbiota and the following process-related or environmental variables were investigated: Digester technology, primary substrate, temperature in the digester, and continent (see definitions above). All variables were treated as factors.

For alpha diversity analyses, samples were rarefied to 10,000 reads, and alpha diversity (observed ASVs and inverse Simpson) was calculated using the ampvis2 package. The Kruskal-Wallis with Dunn's post-hoc test (Bonferroni correction with α=0.01 before correction) was used to determine statistically significant differences in alpha diversity between samples grouped by process and environmental variables.

865 Beta diversity distances based on Bray-Curtis (abundance-based) for genera was
866 calculated using the vegdist function in the vegan R package and visualized by PCoA
867 plots with the ampvis2 package. To determine how much individual parameters
868 affected the structure of the microbial community across the ADs, a permutational
869 multivariate analysis of variance (PERMANOVA) test was performed on the beta-
870 diversity matrices using the adonis function in the vegan package with 999
871 permutations.

872 Core taxa (genera and species) were determined separately for ADs treating different
873 primary substrates and operating at different temperatures (mesophilic and
874 thermophilic) based on their relative abundances in individual ADs according to the
875 three short-read amplicon datasets. Core taxa definitions were identical to those applied
876 in the MiDAS global survey of WWTPs [20]. Taxa were considered abundant when
877 present at >0.1% relative read abundance in individual ADs. Based on how frequently
878 taxa were observed to be abundant, we defined the following core communities: loose
879 core (>20% of ADs), general core (>50% of ADs), and strict core (>80% of ADs).
880 Additionally, we defined conditionally rare or abundant taxa (CRAT) [100] composed of
881 taxa present in one or more ADs at >1% relative abundance, but not belonging to the
882 core taxa.
883
884 **Data Availability:**
885 The raw and assembled sequencing data generated in this study have been deposited in
886 the NCBI SRA database under accession code PRJNA1019951
887 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1019951]. The MiDAS 5 reference
888 database in SINTAX, QIIME and DADA2 format is available at the MiDAS fieldguide
889 website [https://www.midasfieldguide.org/guide/downloads].
890
891 **Code Availability:**
892 R scripts used for data analyses and figures are available at GitHub
893 [https://github.com/msdueholm/MiDAS5] [101]. Raw data files for the R scripts are
894 available at Figshare [https://doi.org/10.6084/m9.figshare.24219199.v2] [102].

References:

1. Tiwary, A., Williams, I. D., Pant, D. C. & Kishore, V. V. N. Emerging perspectives on environmental burden minimisation initiatives from anaerobic digestion technologies for community scale biomass valorisation. *Renewable and Sustainable Energy Reviews* **42**, 883–901 (2015).
2. Achinas, S., Achinas, V., & Gerrit Jan Willem Euverink. A technological overview of biogas production from biowaste. *Engineering* **3**, 299–307 (2017).
3. Samoraj, M. *et al.* The challenges and perspectives for anaerobic digestion of animal waste and fertilizer application of the digestate. *Chemosphere* **295**, 133799 (2022).
4. Czekała, W., Jasiński, T., Grzelak, M., Witaszek, K. & Dach, J. Biogas plant operation: Digestate as the valuable product. *Energies* **15**, 8275 (2022).
5. Briones, A. & Raskin, L. Diversity and dynamics of microbial communities in engineered environments and their implications for process stability. *Current Opinion in Biotechnology* **14**, 270–276 (2003).
6. Dueholm, M. S. *et al.* 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 (2020).
7. Jiang, C. *et al.* Characterizing the growing microorganisms at species level in 46 anaerobic digesters at Danish wastewater treatment plants: A six-year survey on microbial community structure and key drivers. *Water Research* **193**, 116871 (2021).
8. Vanwonterghem, I. *et al.* Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *ISME Journal* **8**, 2015–2028 (2014).
9. Ofiţeru, I. D. *et al.* Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences* **107**, 15345–15350 (2010).
10. Kirkegaard, R. H. *et al.* The impact of immigration on microbial community composition in full-scale anaerobic digesters. *Scientific Reports* **7**, 9343 (2017).
11. Werner, J. J. *et al.* Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4158–4163 (2011).
12. Campanaro, S. *et al.* Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. *Biotechnology for biofuels* **9**, 26 (2016).
13. Calusinska, M. *et al.* A year of monitoring 20 mesophilic full-scale bioreactors reveals the existence of stable but different core microbiomes in bio-waste and wastewater anaerobic digestion systems. *Biotechnology for Biofuels* **11**, 196 (2018).

14. Mei, R. *et al.* Operation-driven heterogeneity and overlooked feed-associated populations in global anaerobic digester microbiome. *Water Research* **124**, 77–84 (2017).
15. Ma, S. *et al.* A microbial gene catalog of anaerobic digestion from full-scale biogas plants. *GigaScience* **10**, giaa164 (2021).
16. Campanaro, S. *et al.* New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. *Biotechnology for Biofuels* **13**, 25 (2020).
17. McIlroy, S. J. *et al.* MiDAS 2.0: An ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. *Database* **2017**, bax016 (2017).
18. Nierychlo, M. *et al.* MiDAS 3: An ecosystem-specific reference database, taxonomy and knowledge platform for activated sludge and anaerobic digesters reveals species-level microbiome composition of activated sludge. *Water Research* **182**, 115955 (2020).
19. McIlroy, S. J. *et al.* MiDAS: The field guide to the microbes of activated sludge. *Database : the journal of biological databases and curation* **2015**, bav062 (2015).
20. Dueholm, M. K. D. *et al.* MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nat Commun* **13**, 1908 (2022).
21. Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H. & Nielsen, P. H. Back to basics - the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PloS one* **10**, e0132783 (2015).
22. Kristensen, J. M., Singleton, C., Clegg, L.-A., Petriglieri, F. & Nielsen, P. H. High diversity and functional potential of undescribed "Acidobacteriota" in Danish wastewater treatment plants. *Front. Microbiol.* **12**, 906 (2021).
23. Nierychlo, M. *et al.* Low global diversity of *Candidatus* Microthrix, a troublesome filamentous organism in full-scale WWTPs. *Front Microbiol* **12**, 690251 (2021).
24. Petriglieri, F. *et al.* "*Candidatus* Dechloromonas phosphoritropha" and "*Ca.* D. phosphorivorans", novel polyphosphate accumulating organisms abundant in wastewater treatment systems. *ISME J* **15**, 3605–3614 (2021).
25. Petriglieri, F. *et al.* Reevaluation of the phylogenetic diversity and global distribution of the genus "*Candidatus* Accumulibacter". *mSystems* **7**, e00016-22 (2022).
26. Singleton, C. M. *et al.* The novel genus, '*Candidatus* Phosphoribacter', previously identified as *Tetrasphaera*, is the dominant polyphosphate accumulating lineage in EBPR wastewater treatment plants worldwide. *ISME J* **16**, 1605–1616 (2022).
27. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* **12**, 635–645 (2014).
28. Rivière, D. *et al.* Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *ISME Journal* **3**, 700–714 (2009).

29. Fujimoto, M., Carey, D. E., Zitomer, D. H. & McNamara, P. J. Syntroph diversity and abundance in anaerobic digestion revealed through a comparative core microbiome approach. *Appl Microbiol Biotechnol* **103**, 6353–6367 (2019).
30. Westerholm, M., Calusinska, M. & Dolfing, J. Syntrophic propionate-oxidizing bacteria in methanogenic systems. *FEMS Microbiology Reviews* **46**, fuab057 (2022).
31. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* **10**, 1–11 (2019).
32. McDonald, D. *et al.* Greengenes2 unifies microbial data in a single reference tree. *Nat Biotechnol* 1–4 (2023) doi:10.1038/s41587-023-01845-1.
33. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* **15**, 796–798 (2018).
34. Bahram, M., Anslan, S., Hildebrand, F., Bork, P. & Tedersoo, L. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. *Environmental Microbiology Reports* **11**, 487–494 (2018).
35. Pausan, M. R. *et al.* Exploring the archaeome: Detection of archaeal signatures in the human body. *Frontiers in Microbiology* **10**, 2796 (2019).
36. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nature Microbiology* **4**, 1183–1195 (2019).
37. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
38. Hu, H. *et al.* Global abundance patterns, diversity, and ecology of Patescibacteria in wastewater treatment plants. *Microbiome* **12**, 55 (2024).
39. Mei, R., Narihiro, T., Nobu, M. K., Kuroda, K. & Liu, W.-T. Evaluating digestion efficiency in full-scale anaerobic digesters by identifying active microbial populations through the lens of microbial activity. *Scientific Reports* **6**, 34090 (2016).
40. Sun, L., Pope, P. B., Eijsink, V. G. H. & Schnürer, A. Characterization of microbial community structure during continuous anaerobic digestion of straw and cow manure. *Microbial Biotechnology* **8**, 815–827 (2015).
41. Karakashev, D., Batstone, D. J. & Angelidaki, I. Influence of environmental conditions on methanogenic compositions in anaerobic biogas reactors. *Applied and Environmental Microbiology* **71**, 331–338 (2005).
42. Levén, L., Eriksson, A. R. B. & Schnürer, A. Effect of process temperature on bacterial and archaeal communities in two methanogenic bioreactors treating organic household waste. *FEMS Microbiology Ecology* **59**, 683–693 (2007).
43. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**, aac9323 (2015).
44. Knights, D. *et al.* Rethinking "enterotypes". *Cell Host & Microbe* **16**, 433–437 (2014).

1018 45. Saunders, A. M., Albertsen, M., Vollertsen, J. & Nielsen, P. H. The activated sludge
1019 ecosystem contains a core community of abundant organisms. *The ISME journal*
1020 **10**, 11–20 (2016).
1021 46. Schorn, S. *et al.* Diverse methylotrophic methanogenic archaea cause high methane
1022 emissions from seagrass meadows. *Proceedings of the National Academy of*
1023 *Sciences* **119**, e2106628119 (2022).
1024 47. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota
1025 revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).
1026 48. Nobu, M. K., Narihiro, T., Kuroda, K., Mei, R. & Liu, W. T. Chasing the elusive
1027 Euryarchaeota class WSA2: genomes reveal a uniquely fastidious methyl-reducing
1028 methanogen. *The ISME Journal* **10**, 2478–2487 (2016).
1029 49. Jung, M.-Y. *et al.* A hydrophobic ammonia-oxidizing archaeon of the
1030 *Nitrosocosmicus* clade isolated from coal tar-contaminated sediment.
1031 *Environmental Microbiology Reports* **8**, 983–992 (2016).
1032 50. Huang, W.-C. *et al.* Comparative genomic analysis reveals metabolic flexibility of
1033 Woesearchaeota. *Nat Commun* **12**, 5281 (2021).
1034 51. Morris, B. E. L., Henneberger, R., Huber, H. & Moissl-Eichinger, C. Microbial
1035 syntrophy: Interaction for the common good. *FEMS Microbiology Reviews* **37**,
1036 384–406 (2013).
1037 52. Hao, L. *et al.* Novel syntrophic bacteria in full-scale anaerobic digesters revealed
1038 by genome-centric metatranscriptomics. *ISME J* **14**, 906–918 (2020).
1039 53. Sieber, J. R., McInerney, M. J. & Gunsalus, R. P. Genomic insights into syntrophy:
1040 the paradigm for anaerobic metabolic cooperation. *Annual Review of Microbiology*
1041 **66**, 429–452 (2012).
1042 54. Westerholm, M., Roos, S. & Schnürer, A. *Syntrophaceticus schinkiigen* nov., sp.
1043 nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from a
1044 mesophilic anaerobic filter. *FEMS Microbiology Letters* **309**, 100–104 (2010).
1045 55. Schnürer, A., Müller, B. & Westerholm, M. *Syntrophaceticus*. in *Bergey's Manual*
1046 *of Systematics of Archaea and Bacteria* 1–10 (John Wiley & Sons, Ltd, 2018).
1047 doi:10.1002/9781118960608.gbm01452.
1048 56. McInerney, M. J., Bryant, M. P., Hespell, R. B. & Costerton, J. W. *Syntrophomonas*
1049 *wolfei* gen. nov. sp. nov., an anaerobic, syntrophic, fatty acid-oxidizing bacterium.
1050 *Applied and Environmental Microbiology* **41**, 1029–1039 (1981).
1051 57. Wu, C., Liu, X. & Dong, X. *Syntrophomonas cellicola* sp. nov., a spore-forming
1052 syntrophic bacterium isolated from a distilled-spirit-fermenting cellar, and
1053 assignment of *Syntrophospora bryantii* to *Syntrophomonas bryantii* comb. nov. *Int*
1054 *J Syst Evol Microbiol* **56**, 2331–2335 (2006).
1055 58. Sousa, D. Z., Smidt, H., Alves, M. M. & Stams, A. J. M. *Syntrophomonas zehnderi*
1056 sp. nov., an anaerobe that degrades long-chain fatty acids in co-culture with
1057 *Methanobacterium formicicum*. *Int J Syst Evol Microbiol* **57**, 609–615 (2007).
1058 59. Hatamoto, M., Imachi, H., Fukayo, S., Ohashi, A. & Harada, H. *Syntrophomonas*
1059 *palmitatica* sp. nov., an anaerobic, syntrophic, long-chain fatty-acid-oxidizing

1060 bacterium isolated from methanogenic sludge. *International Journal of Systematic*
1061 *and Evolutionary Microbiology* **57**, 2137–2142 (2007).
1062 60. Niu, L., Song, L., Liu, X. & Dong, X. *Tepidimicrobium xylanilyticum* sp. nov., an
1063 anaerobic xylanolytic bacterium, and emended description of the genus
1064 *Tepidimicrobium*. *International Journal of Systematic and Evolutionary*
1065 *Microbiology* **59**, 2698–2701 (2009).
1066 61. Wang, G., Li, Q., Gao, X. & Wang, X. C. Sawdust-derived biochar much mitigates
1067 VFAs accumulation and improves microbial activities to enhance methane
1068 production in thermophilic anaerobic digestion. *ACS Sustainable Chem. Eng.* **7**,
1069 2141–2150 (2019).
1070 62. Liu, Y., Balkwill, D. L., Aldrich, H. C., Drake, G. R. & Boone, D. R.
1071 Characterization of the anaerobic propionate-degrading syntrophs *Smithella*
1072 *propionica* gen. nov., sp. nov. and *Syntrophobacter wolinii*. *International Journal*
1073 *of Systematic and Evolutionary Microbiology* **49**, 545–556 (1999).
1074 63. de Bok, F. A. M., Stams, A. J. M., Dijkema, C. & Boone, D. R. Pathway of
1075 propionate oxidation by a syntrophic culture of *Smithella propionica* and
1076 *Methanospirillum hungatei*. *Applied and Environmental Microbiology* **67**, 1800–
1077 1804 (2001).
1078 64. Dolfing, J. Syntrophic propionate oxidation via butyrate: a novel window of
1079 opportunity under methanogenic conditions. *Applied and Environmental*
1080 *Microbiology* **79**, 4515–4516 (2013).
1081 65. Embree, M., Nagarajan, H., Movahedi, N., Chitsaz, H. & Zengler, K. Single-cell
1082 genome and metatranscriptome sequencing reveal metabolic interactions of an
1083 alkane-degrading methanogenic community. *ISME J* **8**, 757–767 (2014).
1084 66. Tan, B., Nesbø, C. & Foght, J. Re-analysis of omics data indicates *Smithella* may
1085 degrade alkanes by addition to fumarate under methanogenic conditions. *ISME*
1086 *Journal* **8**, 2353–2356 (2014).
1087 67. Nobu, M. K. *et al.* The genome of *Syntrophorhabdus aromaticivorans* strain UI
1088 provides new insights for syntrophic aromatic compound metabolism and electron
1089 flow. *Environmental Microbiology* **17**, 4861–4872 (2015).
1090 68. Qiu, Y.-L. *et al. Syntrophorhabdus aromaticivorans* gen. nov., sp. nov., the first
1091 cultured anaerobe capable of degrading phenol to acetate in obligate syntrophic
1092 associations with a hydrogenotrophic methanogen. *Appl Environ Microbiol* **74**,
1093 2051–2058 (2008).
1094 69. McInerney, M. J. *et al.* The genome of *Syntrophus aciditrophicus*: Life at the
1095 thermodynamic limit of microbial growth. *Proceedings of the National Academy of*
1096 *Sciences* **104**, 7600–7605 (2007).
1097 70. Jackson, B. E., Bhupathiraju, V. K., Tanner, R. S., Woese, C. R. & McInerney, M.
1098 J. *Syntrophus aciditrophicus* sp. nov., a new anaerobic bacterium that degrades fatty
1099 acids and benzoate in syntrophic association with hydrogen-using microorganisms.
1100 *Arch Microbiol* **171**, 107–114 (1999).

1101 71. Ganidi, N., Tyrrel, S. & Cartmell, E. Anaerobic digestion foaming causes – A
1102 review. *Bioresource Technology* **100**, 5546–5554 (2009).
1103 72. Duan, J.-L. *et al.* Unraveling anaerobic digestion foaming via association between
1104 bacterial metabolism and variations in microbiota. *ACS EST Eng.* **1**, 978–988
1105 (2021).
1106 73. Jiang, C. *et al.* Identification of microorganisms responsible for foam formation in
1107 mesophilic anaerobic digesters treating surplus activated sludge. *Water Research*
1108 **191**, 116779 (2021).
1109 74. Bovio-Winkler, P., Cabezas, A. & Etchebehere, C. Database mining to unravel the
1110 ecology of the phylum Chloroflexi in methanogenic full scale bioreactors. *Frontiers*
1111 *in Microbiology* **11**, 603234 (2021).
1112 75. McIlroy, S. J. *et al.* Culture-independent analyses reveal novel Anaerolineaceae as
1113 abundant primary fermenters in anaerobic digesters treating waste activated sludge.
1114 *Frontiers in Microbiology* **8**, 1134 (2017).
1115 76. Singleton, C. M. *et al.* Connecting structure to function with the recovery of over
1116 1000 high-quality metagenome-assembled genomes from activated sludge using
1117 long-read sequencing. *Nature Communications* **12**, 2009 (2021).
1118 77. Jiang, F. *et al.* Recovery of metagenome-assembled microbial genomes from a full-
1119 scale biogas plant of food waste by pacific biosciences high-fidelity sequencing.
1120 *Front. Microbiol.* **13**, (2023).
1121 78. Treu, L., Kougias, P. G., Campanaro, S., Bassani, I. & Angelidaki, I. Deeper insight
1122 into the structure of the anaerobic digestion microbial community; The biogas
1123 microbiome database is expanded with 157 new genomes. *Bioresource Technology*
1124 **216**, 260–266 (2016).
1125 79. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique
1126 molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods* **18**, 165–
1127 169 (2021).
1128 80. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers
1129 for classical and next-generation sequencing-based diversity studies. *Nucleic Acids*
1130 *Research* **41**, e1 (2013).
1131 81. Lane, D. J. 16S/23S rRNA sequencing. in *Nucleic Acid Techniques in Bacterial*
1132 *Systematics* (eds. Stackebrandt, E. & Goodfellow, M.) 115–175 (John Wiley and
1133 Sons, Chichester, United Kingdom, 1991). doi:10.1007/s00227-012-2133-0.
1134 82. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: Assessing
1135 small subunit rRNA primers for marine microbiomes with mock communities, time
1136 series and global field samples. *Environmental Microbiology* **18**, 1403–1414
1137 (2016).
1138 83. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the
1139 Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**, 1621–1624 (2012).
1140 84. Pinto, A. J. & Raskin, L. PCR biases distort bacterial and archaeal community
1141 structure in pyrosequencing datasets. *PLoS One* **7**, e43093 (2012).

85. Apprill, A., Mcnally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* **75**, 129–137 (2015).
86. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
87. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 81257 (2016) doi:10.1101/081257.
88. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
89. Walters, W. A. *et al.* PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**, 1159–1161 (2011).
90. R Development Core Team, R. F. F. S. C. & R Core Team. *R: A Language and Environment for Statistical Computing*. *Vienna Austria R Foundation for Statistical Computing* vol. 1 (R Foundation for Statistical Computing, Vienna, Vienna, 2008).
91. RStudio Team. *RStudio: Integrated Development Environment for R*. (RStudio, PBC., Boston, MA, 2020).
92. Oksanen, J. *et al. Vegan: Community Ecology Package*. (2019).
93. Richard A. Becker, Allan R. Wilks, Ray Brownrigg, Thomas P Minka, & Alex Deckmyn. maps: Draw Geographical Maps. (2021).
94. Dowle, M. & Srinivasan, A. data.table: Extension of `data.frame`. (2019).
95. Ogle, D. H., Doll, J. C., Wheeler, A. P. & Dinno, A. FSA: Simple fisheries stock assessment methods. (2023).
96. Mangiafico, S. rcompanion: Functions to support extension education program evaluation. (2023).
97. Pedersen, T. L. Patchwork: the composer of plots. (2020).
98. Constantin Ahlmann-Eltze. ggupset: Combination matrix axis for 'ggplot2' to create 'UpSet' plots. (2020).
99. Andersen, K. S. S., Kirkegaard, R. H., Karst, S. M. & Albertsen, M. Ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* 299537 (2018) doi:10.1101/299537.
100. Dai, T. *et al.* Identifying the key taxonomic categories that characterize microbial community diversity using full-scale classification: a case study of microbial communities in the sediments of Hangzhou Bay. *FEMS Microbiology Ecology* **92**, fiw150 (2016).
101. Dueholm, M. K. D. msdueholm/MiDAS5: R-scripts for MiDAS 5: Global diversity of bacteria and archaea in anaerobic digesters. Zenodo https://doi.org/10.5281/zenodo.10982338 (2024).
102. Dueholm, M. MiDAS 5: Global diversity of bacteria and archaea in anaerobic digesters (data for R-scripts). (2024) doi:10.6084/m9.figshare.24219199.v2.

Acknowledgements:
The project has been funded by the Danish Research Council (grant 6111-00617A, P.H.N.) and the Villum Foundation (Dark Matter and grant 13351, P.H.N.). We thank all the involved anaerobic digester plants for providing samples and plant metadata.

Author Contributions:
P.H.N. and M.K.D.D. designed the study. M.K.D.D. and P.H.N. wrote the manuscript and all authors reviewed and approved the final manuscript. M.A., Y.B-F., D.B., C.B., M.C.C, Å.D., L.E., C.H., K.K., N.K., C.L., G.L., S.M., V.O., P.O-P., D.P., V.R., M.R., J.R., P.E.S., N.T., J.V., J.D.V., C.W. provided samples and metadata. V.R. handled sampling, DNA extraction and library preparation for DNA sequencing. K.S.A. and M.K.D.D. performed the bioinformatics analyses. M.K.D.D., A-K.C.P. and K.S.A. curated metadata and carried out statistical analyses.

Competing Interests:
The authors declare no competing interests.