ARC Training Centre for the Transformation of Australia's Biosolids Resource

PROGRESS REPORT

Results from a sequencing trial and recommendations for project delivery

Development of a digester-health test using DNA sequencing technology

Project activity period: August 2024 to June 2025
Lead Institution: The Royal Melbourne Institute of Technology (RMIT)
Industry partner: Melbourne Water
Investigators: Christian Krohn, Andy S. Ball, Damien Batstone

Contact details:

Dr. Christian Krohn, christian.krohn@rmit.edu.au - Industry Centre Postdoc

Prof Andy S. Ball, andy.ball@rmit.edu.au - Centre Director

Prof Damien J. Batstone, d.batstone@uq.edu.au - Theme Leader

ARC Training Centre for the Transformation of Australia's Biosolids Resource,
RMIT University, Building 215, Level 3, Room 003-06, RMIT Bundoora West Campus, 225-245 Plenty Road, Bundoora, Victoria 3083, Australia

Project objective:

To develop a DNA-based predictive test based on taxonomic profiles for healthy and efficient digestion that can be used to monitor the health of the ETP (and eventually WTP) digesters for real-time biomass process monitoring. The test should be fast so it can be applied as frequently as possible.

Scope and background

As set out in the project proposal from 22/4/2024.

The work for this report was supported by funding from the RMIT Innovation Proof of Concept fund (AUD$10k) to cover consumables.

Table of Content

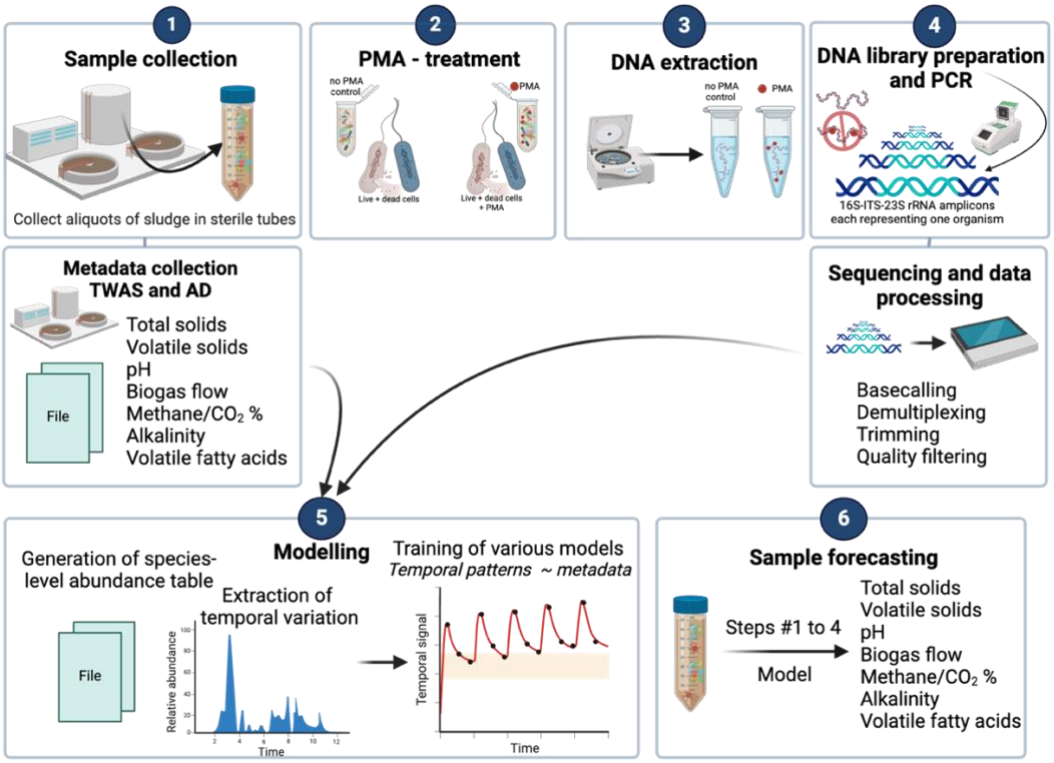

Figure 1. Workflow activities involved in this project.

Definitions:

16S rRNA gene:	The 16S rRNA gene in bacteria and archaea is part of the rRNA operon, encoding the 16S rRNA subunit. It is approximately 1,500 base pairs long. The sequences of this gene are used as markers for microbial identification because they contain various regions with sequences that are species-specific. While these sequences can distinguish between different species, there may not be enough variability to distinguish between all species and they are not sufficient to identify cells at the strain level.
23S rRNA gene:	The 23S rRNA gene in bacteria and archaea is part of the rRNA operon and encodes the 23S rRNA subunit. It is approximately 2,904 base pairs long. These sequences are not commonly used as markers for microbial identification.
16S-ITS-23S:	A DNA sequence in bacteria and archaea that includes most of the rRNA operon. The full DNA sequence, or the individual regions (16S, ITS, 23S), or small regions within each loci, are used for microbial identification.
ITS:	Internal transcribed spacer. Spacer DNA that sits between the 16S and 23S genes on the rRNA operon. Its sequences can be used as markers for microbial identification. However, it is variable in length and sequences, which can vary even between different operon copies within the same genome.
Microbial strain:	A genetic variant or subtype of a microbial species.
Microbial species:	Microbes that share a high degree of genetic and physiological similarity. Species are often determined through genetic sequences, such as from 16S rRNA genes for bacteria.
MiDAS:	Database and field guide for microbes (and their amplicon reference sequences) in activated sludge and anaerobic digesters.
minimap2	A software to map DNA or RNA sequences to reference sequences. It is optimised for Oxford Nanopore reads.
Kraken2	A taxonomic classification software that assigns DNA or RNA sequences to known organisms by comparing them to a database of genomic sequences.
RefSeq:	Database built by the National Center for Biotechnology Information (NCBI) containing reference sequences of genes, proteins and transcripts.
Ribosomal RNA (rRNA):	Non-coding RNA present in all cells. This RNA structurally facilitates protein synthesis (as part of so-called macromolecular machines) and makes up the majority of all cellular RNA. It consists of different structural subunits that include 16S rRNA and 23S rRNA in bacteria and archaea. The genes for producing these subunits are located on the same gene operon and separated by spacer DNA, the internal transcribed spacer (ITS) sequences.
rRNA operon:	A cluster of genes that together encode ribosomal RNA. It contains the 16S, ITS and 23S genes. There are 7–15 copies of the rRNA operon present in a microbial genome.

1. Introduction

Sludge DNA is a rich source of information that is currently underutilised for process diagnostics and monitoring. DNA-based monitoring has applications across the wastewater treatment chain, including improving process stability and enhancing gas recovery during anaerobic digestion processes.

We propose an approach for linking DNA to operational conditions using predictive models that link temporal changes of microbial compositions (derived from DNA sequencing) to changes in process endpoints such as biogas flow and quality. For example, relative changes of methanogens in sludge may predict quantitative changes in methane concentrations. By sampling, sequencing and modelling anaerobic digestate from two full-scale digesters at Melbourne Water over 83 days from June to December 2024, this project will provide a proof of concept for the later development of this 'digester-health test'.

The most common sequencing method used to get semi-quantitative microbial compositions is called 'amplicon sequencing'. Amplicon sequencing relies on a PCR step using specific primers that amplify marker genes, which represent target microorganisms (activity 4 in Fig. 1). The DNA of those amplified marker genes (usually a certain region in the 16S ribosomal RNA gene) is then sequenced to get the sludge composition of bacteria and archaea. For the proposed model approach to work, it is important that the primers cover as many prokaryote microorganisms as possible. Our previous work has shown that currently used primers (V4 and V3-V4) provide robust coverage of process relevant microbial genera (Krohn et al., 2024).

However, the resolution of microbial detection is also important. While the mentioned primers cover a wide range of bacteria and archaea, they can only confidently link microbial identity to genus level, not at species or strain-level. This is mainly due to the length limitations of Illumina sequencing technology, which can only sequence short reads (~150–400 base pairs). Hence, we proposed the use novel Oxford Nanopore (ONP) sequencing technology that can read longer DNA strands compared to the current Illumina instruments (Wang et al., 2021). Longer DNA strands enable the sequencing of the complete ribosomal RNA operon (~4,500 base pairs), rather than only a small region of the gene, and therefore identify microbes to species or even strain level. An important caveat of ONP long-read sequencing is read quality, which is further discussed in Section 3.

We considered the ONP long-read sequencing approach for the predictive models for this project, as higher taxonomic resolution (due to long reads) may provide more nuanced process relevant information for modelling. Additionally, such high taxonomic resolution, once available to industry, will have wide applications for process optimisations, monitoring and precision diagnostics, as a wide range of process relevant species or disease vectors will be detectable. The small ONP devices, which plug into a laptop for running sequencing library, are portable, require little upfront capital expenditure and can be integrated into on-site laboratory monitoring regime.

Hence, a protocol was developed using a novel primer for sequencing the whole ribosomal RNA operon (16S-ITS-23S primers, Fig. 1 activity 4) on anaerobic sludge and the coverage of bacteria and archaea was compared with short-read Illumina sequencing (V3-V4 primers). We herewith report findings of this sequencing trial, describe the technology as well as its current challenges, and provide recommendations for research and development for the digester-health test.

2. Methods

Amplicon DNA of twenty-four anaerobic sludge samples from six anaerobic laboratory-scale (5 L) RMIT reactors were sequenced after amplifying genes that encode the 16S-ITS-23S rRNA operon. A protocol for library preparation and amplicon sequencing was developed and was made available online: https://chris-krohn.github.io/ABlab-workflows-longreads.

About 50 fmol of DNA library containing a pool of amplicons (4.25 kb length) for 24 samples were loaded onto a MinION flow cell (R10.4.1, #FAZ29172, 1,525 available pores), which was run for 24 hours and provided 75 Gb of pod5 data. The pod5 data was basecalled at Super High Accuracy (SUP) with dorado (v0.8.2) into 4.85 Gb of fastq files with a total of 576,790 amplicon reads (median read length 4,109 bp at median Q24 quality score). A higher data yield was expected, and it was noted that additional 36.5 Gb of pod5 data was lost due to a power outage during the sequencing processes and was not included in the analysis.

2.1. Taxonomic classification

The amplicon reads were quality trimmed using `chopper` (De Coster and Rademakers, 2023) as described in the online protocol (min quality Q20, min and max read length 3,000 and 5,000 respectively). Different methods were subsequently used to classify taxonomy of the filtered amplicons. Generally, there are two types of classification methods available and were evaluated for the 16S-ITS-23S amplicons: (1) kmer-based classification and (2) alignment-based classification, which are used for classification of metagenomic and amplicon reads (Wright et al., 2023).

First, a kmer-based method was employed in this study, using the 'wf-metagenomics' workflow (v2.10.1), available through Nanopore's Epi2Me platform (https://github.com/epi2me-labs/wf-metagenomics), employing Kraken2 in conjunction with Bracken (Lu et al., 2022; Wood et al., 2019) with the NCBI RefSeq target loci database (16S, 18S, 28S, ITS) covering Bacteria, Archaea and Fungi (www.ncbi.nlm.nih.gov/refseq/targetedloci). K-mers are short, overlapping subsequences of length k (35 bp by default) and are mapped via exact-matching of read k-mers against a Kraken database of k-mers from existing genomes or target loci sequences. The Kraken2 ` --confidence` threshold was set to 0 as per Epi2Me default. This parameter requires careful calibration with known standards, which was out of the scope for this study (Wright et al., 2023).

To compare k-mer based methods with read alignment methods, the same Epi2Me platform was chosen to run the alignment-based workflow, which employs minimap2 (Li, 2021, 2018) and samtools with following settings: `minimap2 --cap-kalloc 100m --cap-sw-mem 50m -ax map-ont` and `samtools view -h -F 2304` (-F 2304 flag removes secondary and supplementary alignments), also with the RefSeq target loci database.

Using the wf-metagenomics workflow for both, Kraken2 and minimap2 had the advantage that unclassified reads were quantified, reported and compared. However, we note that these tools currently do not provide a confidence score for classification accuracy.

2.2. Comparison of GROND 16S-ITS-23S with MiDAS 16S database

While the RefSeq target loci database contains curated, non-redundant reference sequences for different RNA loci (full length 16S, 18S, 28S, ITS), it does not contain reference sequences of full RNA operons (16S-ITS-23S rDNA), with potential implications for classification accuracy. The use of 16S-ITS-23S databases provides classification resolution of more reads to species level, and even strain-level. Presently there are only two whole operon databases available, GROND and MIrROR (Seol et al., 2022; Walsh et al., 2024). We opted for the GROND database as it contained a higher number of reference sequences. We used a custom mapping protocol with minimap2 (minimap2 -K20M -t 13 --secondary=no` and `samtools view -F 4 -F 256 -F 2048` to remove unmapped, secondary and supplementary alignments) to map the samples reads to the GROND reference sequences (Walsh et al., 2024).

Furthermore, the same custom minmap2 protocol was utilised to map the reads to the wastewater-specific MiDAS53 database (MiDAS version 5.3), which contains full-length 16S references sequences derived from sludges across the globe. While its reference sequences only cover the 16S rRNA gene (and do not include the additional ITS or 23S loci), MiDAS53 is currently the most comprehensive wastewater database, also enabling species-level classification, specific to engineered wastewater environments.

We further note that the scope of this experiment did not include the evaluation of classification performance as no community mock standards were included in the experimental design. Instead, we aimed to gain a general understanding of suitability and bias of classification methods and databases for 16S-ITS-23S sequencing from wastewater sludge, compared to more commonly used methods.

2.3. Comparison to conventional 16S profiles

The genus-level taxonomic composition based on 16S-ITS-23S rRNA operons was compared to the taxonomic composition from conventional primers (V3-V4) and to the taxonomic composition of metagenome-extracted full-length 16S rDNA sequences. Here, the reference data, which was based on V3-V4 primers and metagenomes, were produced from the same DNA that was used for this experiment and were made available by the ARC Biosolids Training Centre. This reference data was generated as

part of activities that involved the operation of six continuously stirred-tank reactors (CSTRs) for 23 weeks. The amplicon read abundances and classifications based on V3-V4 primers and metagenome-extracted full-length 16S rDNA sequences were set as the reference for the comparisons with the 16S-ITS-23S reads in this review. More information about the reference data is available in the manuscript by Krohn et al., (2025) ("*Ecology of foam establishment during anaerobic digestion of wastewater sludge following process disturbance*" Water Research, under review).

To account for different sequencing depths, each set of classified reads was normalised to lowest number of sample reads via random subsampling without replacement (rarefication) before comparing abundance profiles, using the `vegan` package (Oksanen et al., 2008).

3. Results and discussion

Sequencing 16S-ITS-23S amplicons with a MinION flow cell enabled species and strain-level detection of bacteria and archaea in anaerobic sludge (Table 1). The initial total consumables expenditure for consumables was AUD $8,000, which was estimated to be ~AUD $92.5 per sample, including those listed in Table 2, if all consumables were maximally utilised.

Table 1. Overview of methods and databases used for classification of 16S-ITS-23S amplicons for this study, and total number of species classified.

Classification method	Functionality	Database	Species classified	Strains classified
Kraken2, v2.1.2	Exact k-mer matching to reference sequences	RefSeq target loci 16S, 18S, 28S, ITS	1,399	0
Minimap2, v2.24-r1122	Read alignment to reference sequences, Smith–Waterman algorithm, long-read optimised.		1,102	0
		MiDAS v.5.3 16S	547	0
		GROND 16S-ITS-23S	338	57

Table 2. Estimated consumables costs per sample for sequencing 16S-ITS-23S amplicons of 24 sludge samples, including DNA extraction, PCR, library preparation and sequencing on an Oxford Nanopore MinION flow cell.

Consumables	Supplier	Costs / sample (AUD $)
Flow cell R10.4.1	ONP	45
Native Barcoding kit SQK-NBD114.96	ONP	13
DNeasy Powersoil Pro #47016	Qiagen	9
Consumables for PCR and library preparation (Various: see here)	New England Biolab	22.5
Other: for example, magnetic beads for PCR clean up, Qubit DNA quantification	Various	3
Total		92.5

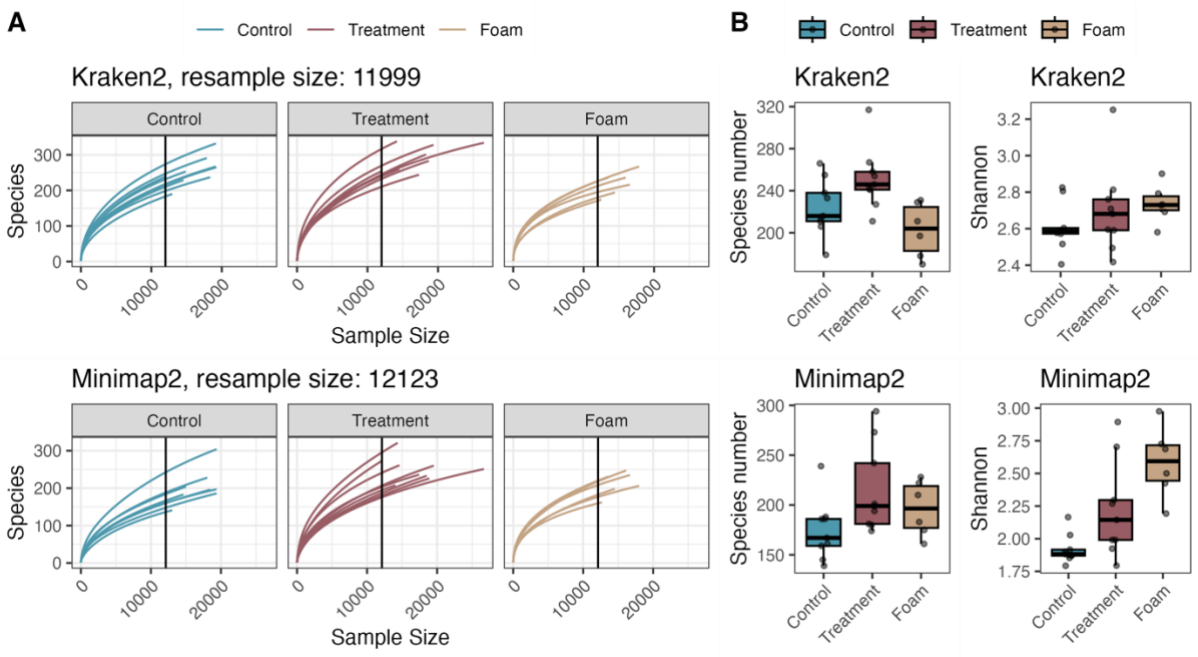

Figure 2. A. Comparison of rarefaction curves between two different classification methods, Kraken2 and minimap2. Abundance tables of both, Kraken2 and minimap2 workflows, were generated with the Epi2Me (both based on the Refseq target loci database). The vertical line indicates the lowest common sample size (resample size used for rarefication). The lines show increasing numbers of species; however, they were not covered in sufficient depth as the lines have not plateaued with increasing sample size. It therefore indicated that a greater sequencing depth (i.e. longer sequencing duration) was required to capture species diversity sufficiently. **B.** Boxplots of species numbers and Shannon diversity indices, calculated after normalisation (rarefication). It showed that the two shown classification methods did result in a different species diversity. The control, treatment and foam categories are not relevant for this study and used as reference.

3.1. Read depth

Overall, it was apparent that read depth after 24 hours of sequencing starting with 1,525 active pores was not sufficient. As illustrated in the rarefaction curves in Figure 2A, the read depth after acquiring data during 24 hours of sequencing did not cover species diversity sufficiently from anaerobic sludge. After quality trimming with `chopper' and classification with minimap2 and Kraken2 on the Epi2Me platform, a sample median of 16,268 and 16,416 reads were obtained, respectively.

A minimum of 100,000 reads per samples is often recommended to cover diversity appropriately for ecological assessments, hence more data is required for future sequencing runs. Accounting for the around 33% data loss due to a power outage during this reported sequencing run, the MinION flow cell acquired a total of around 24,000 quality filtered reads per hour. Hence, it is recommended that a library of 24 samples, loaded on a MinION flow cell at 50 fmol concentration, is sequenced until all flow cell pores are spent (around 72 hours) to maximise diversity coverage. A final yield of 24,000 quality filtered reads per hour equates to an estimated total of 1,728,000 reads after 72 hours, which is significantly less compared to the theoretical total number reads for a MinION flow cell (Table 3). This was partly

due to downstream filtering of reads not passing quality/length requirements, and likely further due to sub-optimal library quality and flow cell loading.

Table 3. Comparison of theoretical sequence yields (nucleotide bases and read counts) of three available Nanopore flow cells, based on assumed average active channels. The MinION FLO-MIN114 was used in this study.

Flow cell	Channels	Speed (bases/sec/ channel)	Run time (hours)	Theoretical yield[a] (Gigabases)	Theoretical reads[b] (counts)
Flongle FLO-FLG114	126	50–400	16	1.45	0.34M
MinION FLO-MIN114	512	50–400	72	26.5	6.46M
PromethION FLO PRO114M	2,675	200–500	72	208	50.60M

[a]Assuming 200 (Flongle/MinION) and 300 (PromethION) active channels on average; variable depending on active pores and DNA library quality, [b]assuming an average read length of 4,109bp, M; million

3.2. Read quality

Mean read quality of this sequencing trial was around Q24.6 (~99.65% accuracy) after quality filtering, slightly lower than the expected base calling quality of Q26 (99.75% accuracy) with R10.41 pore chemistry according to ONP. While ONP read quality is steadily improving it remains lower compared to the gold standard from Illumina (Q30-Q50). Read quality of ONP sequencing technology is rapidly improving, since the first portable devices were released by Oxford Nanopore in 2014 (Bayley, 2015; Deamer et al., 2016; Rang et al., 2018; Wang et al., 2021). Given the consistent improvements in pore engineering and signal detection algorithms it is anticipated that base calling accuracy will surpass Q30 (99.90%) from 2025 onwards.

At Q26 average base calling accuracy, 2.5 bases per each 1000 bases are falsely called by the device (0.25% error rate), or 10 errors in each of 4,000bp amplicons. This would result in the removal of most of the 4,109bp-long reads during subsequent data processing steps. For example, popular tools such as DADA2 (Callahan et al., 2016) are designed to call and quantify each sequence variant of the amplicon reads and remove reads with more than 2 errors (although accepted error rates can be adjusted), creating accurate, 'denoised' abundance tables of amplicon sequence variants.

Similarly, common tools that create abundance tables after clustering of amplicon reads into operational taxonomic units (OTUs) are sensitive to nanopore errors. Systematic base call errors are propagated from nanopore signal translation of certain gene regions, such as homopolymers (Rang et al., 2018) and may result in spurious clusters. Bespoke clustering tools for long amplicon reads are presently developed that address such ONP specific errors, by integrating various polishing tools such as Racon and Medeka (Rodríguez-Pérez et al., 2021) or by developing specific algorithms that iteratively adapt classification probabilities based on profile estimates (Curry et al., 2022).

However, due to the challenges associated with ONP base calling errors, some studies prefer to forgo any read denoising or clustering steps. As done for this study, reads can be directly aligned to a database using more general read mapping software, such as minimap2 (Li, 2021, 2018), or directly mapped using k-mers as done with Kraken2/Bracken (Lu et al., 2022; Wood et al., 2019). That is simpler but comes at a cost of lower representation of ground truth in terms of abundances.

3.3. Comparison of two classification methods

Using Oxford Nanopore's `wf-metagenomics` workflow from the Epi2Me platform with default settings enabled the comparison of two established classification methods to classify the long 16S-ITS-23S amplicons, while controlling for other sources of variability. The curated RefSeq Targed Loci database was used for this purpose as it was accessible in the `wf-metagenomics` workflow. Unclassified sequences were included in the final abundance tables. A lower number of reads were classified with minimap2 compared to Kraken2 (Figs. 2B and 3). In total 24 and 35 and phyla were classified with minimap2 and Kraken2, respectively (Figs. 3 and 4). Only 35 % of the detected species were identical between the two classification methods.

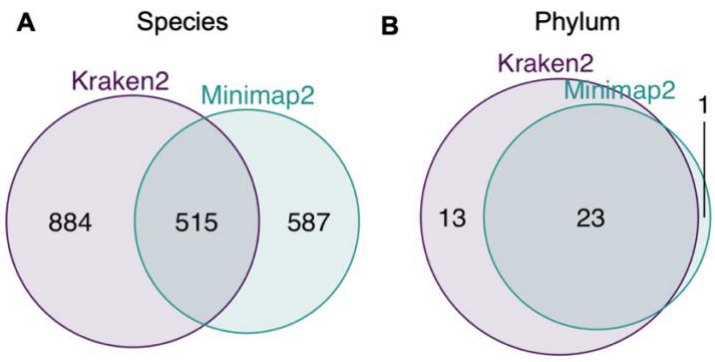

Figure 3. Number (**A**) species and (**B**) phyla - detected with two different classification methods, Kraken2 and minimap2. Abundance tables of both, Kraken2 and minimap2 workflows, were generated with Epi2Me using the Refseq target loci database.

Proteobacteria were detected with Kraken2 and made up a large fraction of total reads with up to 18% of relative abundances but remained undetected with minimap2 (Fig. 4). The majority of Proteobacteria abundances with Kraken2 was represented by two species, *Acidomonas methanolica* and *Rodentibacter rarus* (5.2% and 2.7% of total reads respectively), which remained unclassified with minimap2. Proteobacteria were also missing when using minimap2 with MiDAS53 and the 16S-ITS-23S database, indicating that this difference was driven by the classification method (Figs. 5 and 6). Other taxa also remained unclassified with minimap2, compared to Kraken2, however all remaining unknowns were of low abundance.

It is possible that these additionally detected species with Kraken2 represent false positives. Kraken2 is said to have high precision (high ratio of true positives to false positives), while its sensitivity or recall

is known to be low (meaning there are high numbers of false negatives or non-detects), depending on database size (Wood et al., 2019; Wright et al., 2023). Consequently, those that are classified are likely correct provided that the confidence threshold is set appropriately. Confidence thresholds in Kraken2 are calculated based on the proportion of the number of k-mers that are unambiguously mapped to a taxon. Setting higher (>0) confidence thresholds is meant to improve precision at the cost of recall (Wright et al., 2023). However, there is little guidance on the appropriate use of this confidence threshold in Kraken2 for targeted loci sequences. Furthermore, confidence scores that would inform classification performance for each taxon are currently not reported (Wood et al., 2019; Wright et al., 2023). Hence, it was decided to keep the confidence threshold at 0 in this experiment as per default.

Nonetheless, in agreement with Kraken2, the distribution of *A. methanolica* abundances in the context of the experimental samples used (the control and treatment) aligned with the functional potentials of this species. The treated samples used for this sequencing trial included inhibited digestate with elevated alcohols levels, including methanol (Krohn et al. 2025, under review). *A. methanolica* is a facultative methylotroph that oxidises methanol and its relative abundances were ~3 times greater compared to the control, which agreed with elevated methanol levels. On the other hand, the detection of species *R. rarus*, a very rare microaerophilic acidogen and animal pathogen, although possible, is less likely a part of anaerobic digesters (Adhikary et al., 2017). Hence, it remains unclear whether Kraken2 has successfully detected more taxa from the reads or whether they are false positives.

Recommendations for ongoing validation of this work are provided in Section 4.

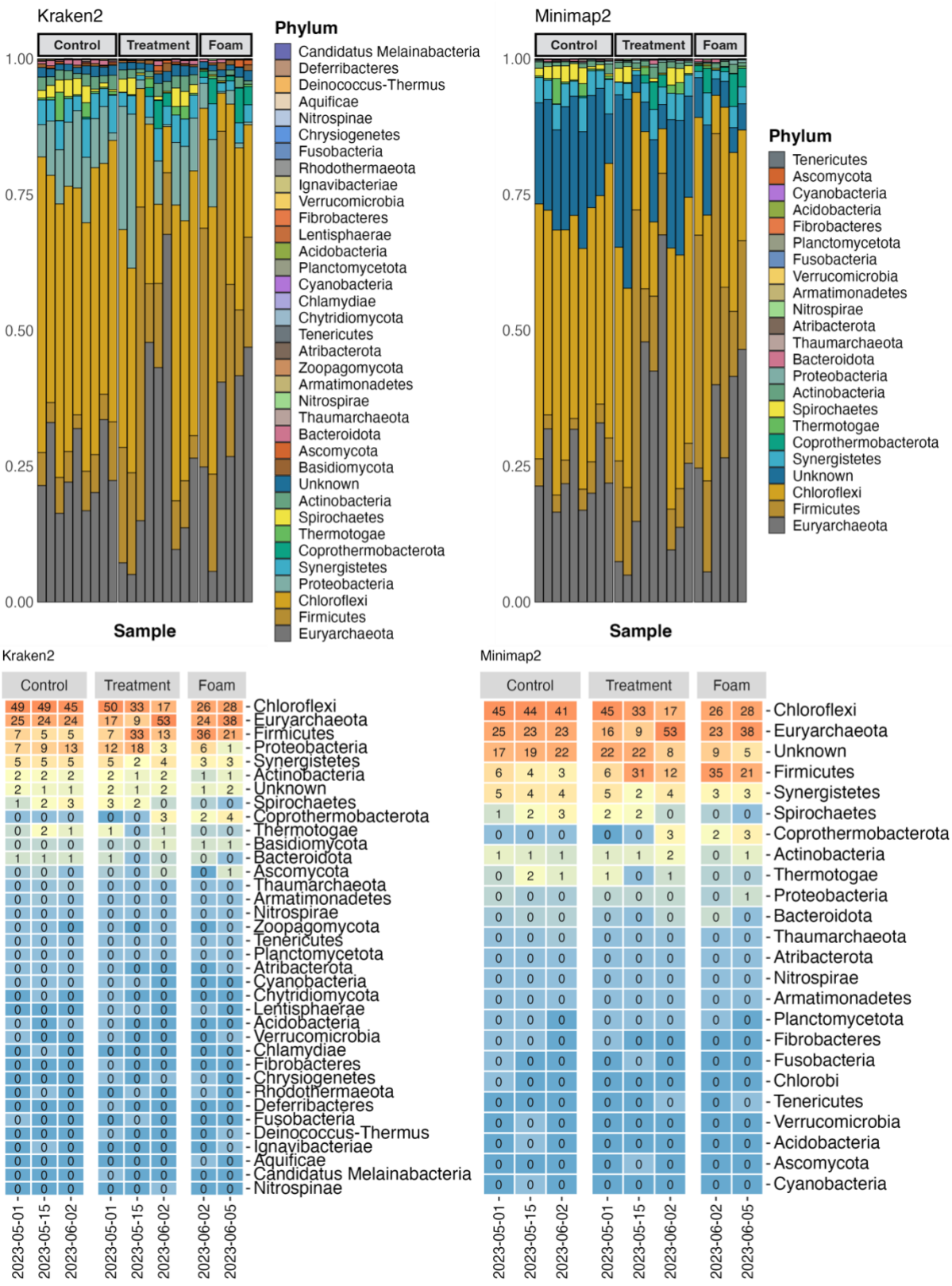

Figure 4. Comparison of relative abundances between two different classification methods, Kraken2 and minimap2. Abundance tables of both, Kraken2 and minimap2 workflows, were generated with Epi2Me using the Refseq target loci database. **Top**: Barplots of 24 samples, identifying 35 (Kraken2) and 23 phyla (minimap2). **Bottom**: Heatmap of the same data but in average abundances of three replicates per day of sampling (in % per day). The control, treatment and foam categories are not relevant for this analysis and used as a reference.

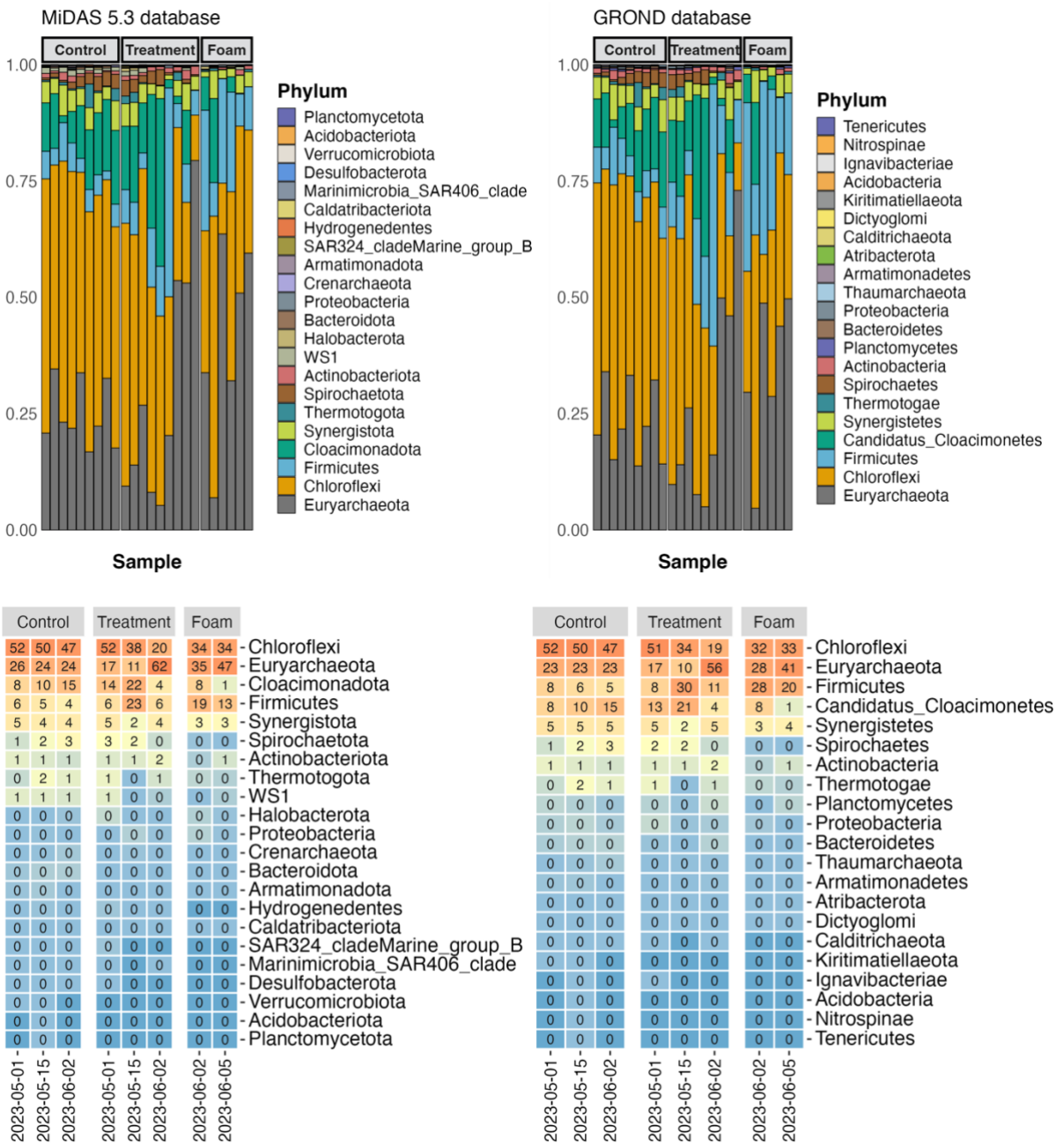

Figure 5. Comparison of relative abundances of 16S-23S-ITS amplicons aggregated to **phylum** level, classified with the wastewater reference database MiDAS53 and the GROND whole operon database (RefSeq version). Sample sequences were mapped to the database sequences with minimap2. **Top**: Barplots of 24 samples, identifying 22 (MiDAS53) and 21 phyla (GROND). **Bottom**: Heatmap of the same data but in average abundances of three replicates per day of sampling (in % per day). The control, treatment and foam categories are not relevant for this analysis and used as a reference.

3.4. Choice of database for species level detection

We used the minimap2 approach to qualitatively compare taxonomic classification performance between the GROND (16S-ITS-23S) and MiDAS53 (16S) databases. As GROND contains whole operon reference sequences (spanning across 16S, ITS and 23S loci with ~4,500 bp) it was expected that it resulted in improved species detection compared to MiDAS53, which contains 16S reference

sequences (~1,500 bp). It was found that GROND indeed outperformed species detection compared to the MiDAS53.

At phylum-level there was little difference between taxonomic profiles of the two databases (Fig. 5), showing agreement in higher order phylogenetic assignments. However, GROND enabled detection of species that MiDAS53 did not. For example, with GROND it was possible to detect and quantify *Candidatus* Syntrophosphaera thermopropionivorans and *Candidatus* Cloacimonas acidaminovorans, which remained unclassified with MiDAS53 (Fig. 6). *Ca.* S. thermopropionivorans is a key microbe for anaerobic digestions as it is one of few known syntrophic propionate-oxidizing bacteria (SPOB), able to utilise propionate and promote its conversion to methane (Dyksma and Gallert, 2019). Propionate degradation is an important rate-limiting step for anaerobic digestion with crucial knowledge gaps in the role of SPOBs.

The phylogenetic identity of *Ca.* S. thermopropionivorans based on GROND was verified with metagenome assembled genomes of the same sludge (Krohn et al., 2025, under review). Furthermore, the increased relative abundances of *Ca.* S. thermopropionivorans on the 15 May 2023 in this study agreed with increased propionate concentrations in digestate (Fig. 6) (Krohn et al., 2025, under review). The genome of *Ca.* S. thermopropionivorans is very similar to the uncultured *Cloacimonadaceae* W5 (Dyksma and Gallert, 2019), which is likely the reason why it was classified as such with MiDAS53 (Fig. 6).

Similarly, the presence at high abundances of *Ca.* Cloacimonas acidaminovorans, a difficult-to-culture syntrophic bacterium (Pelletier et al., 2008), indicated protein degradation and the fermentation of amino acids. This microbe disappeared from the treated digesters from the 15th of May 2023, likely due to sludge acidification, with its disappearance therefore clearly indicating a significant reduction in digestion performance.

Another interesting microbe that was detected and enriched in acidified sludge was *Romboutsia ilealis* (Fig. 6). *R. ilealis*, a known probiotic intestinal species, is capable of producing extracellular soluble fibers (Chang et al., 2023). These fibers ((1,3;1,4)-β-D-glucans), may help this species to physically protect their cell walls under acidic conditions. As such, polysaccharide fibres can make sludge more viscous - hence could play an important role in sludge foaming or bulking phenomena. However, the identify of *R. ilealis* in our study requires validation as it was not clearly detectable from metagenome assembled genomes.

Furthermore, *Clostridiodes difficile* was detected and enriched in the acidified, inhibited digesters (Fig. 6), indicating poor pathogen removal during the time of process inhibition. *C. difficile* is an important emerging human pathogen according to the Centers for Disease Control and Prevention (CDC).

These examples highlighted the relevance of species-level detection for microbial profiling of wastewater treatment processes. The detection of species-level microbial shifts allows operators to

identify, and trouble shoot potential process inhibition events. Hence, the use of a technology that enables reliable detection of as many species as possible is highly desirable for sludge monitoring.

Table 4. Strains detected with 16S-ITS-23S sequencing (GROND database). Thirteen out a total of 57 strains are shown.

Strain	Mean prevalence (out of 24 samples)	Total abundance	Relative abundance (%)
Candidatus Cloacimonas acidaminovorans Evry	21.5	12,668	3.4
Acetomicrobium mobile DSM 13181	23	6,557	1.8
Treponema caldarium DSM 7334	20	5,369	1.5
Anaerolinea thermophila UNI-1	14.5	3,341	0.92
Methanothermobacter thermautotrophicus Delta H	14	2,363	0.65
Fervidobacterium nodosum Rt17-B1	23	1,538	0.42
Syntrophomonas wolfei subsp. wolfei Goettingen G311	20	1,440	0.40
Acetivibrio thermocellus DSM 2360	18	1,268	0.35
Acetivibrio thermocellus ATCC 27405	15.3	910	0.25
Acetivibrio clariflavus 4-2a	17.7	682	0.19
Acetivibrio clariflavus DSM 19732	13	637	0.18
Syntrophomonas zehnderi OL-4	23	609	0.17
Methanosarcina thermophila CHTI-55	18	418	0.11

3.5. Strain-level detection

An advantage of whole operon sequencing is the ability to profile a greater proportion of microbes at species and even strain level, potentially linking genomic information and their economic impact in wastewater treatment. However, strain-level detection of 16S-ITS-23S amplicons requires a database with 16S-ITS-23S reference sequences that is annotated to strain level.

By mapping the 16S-ITS-23S amplicons of anaerobic sludge samples in this study to the GROND 16S-ITS-23S database (Walsh et al., 2024) with minimap2, 57 microbes were detected on strain-level (Table 4). Three of the most abundant strains identified (> 1% relative abundance) were *Candidatus* Cloacimonas acidaminovorans strain Evry (Pelletier et al., 2008), *Acetomicrobium mobile* DSM 13181 (formerly *Anaerobaculum mobile*) (Mavromatis et al., 2013) and *Treponema caldarium* DSM 7334 (Brune et al., 2022; Pohlschroeder et al., 1994).

T. caldarium (known as *Gracilinema caldarium* gen. nov. since Brune et al., 2022) has a fermentative metabolism but was associated with increased cellulose degradation in combination with cellulolytic bacteria (Pohlschroeder et al., 1994). Growth of *Acetomicrobium mobile* DSM 13181 was enhanced in co-culture with *Methanothermobacter thermautotrophicus*) (Mavromatis et al., 2013), which was also present in the digesters (Fig. 6).

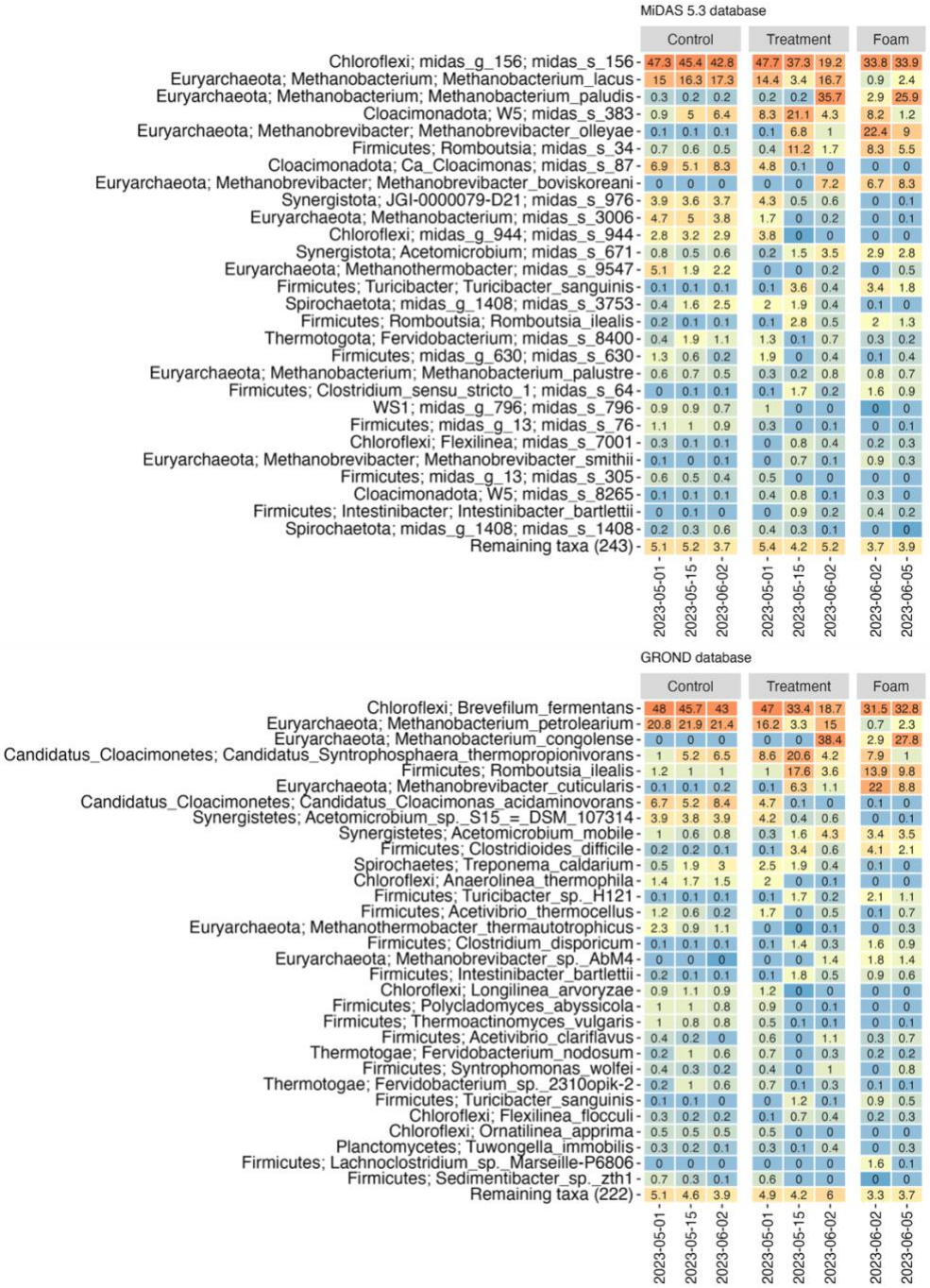

Figure 6. Comparison of relative abundances of 16S-23S-ITS amplicons aggregated to species level, classified with the wastewater reference database MiDAS53 and the GROND whole operon database (RefSeq version). Sample sequences were mapped to the database sequences with minimap2. Heatmap shows average relative abundances (% per day) of the most abundant species, representing 95% of total abundances. All data was rarefied (normalised). Some sequences were given a 'midas' placeholder if identify is not yet known. The control, treatment and foam categories are not relevant for this analysis and used as a reference.

3.6. Primer bias and taxonomic coverage

The primer chosen for this study (A519F/ U2428R, approximately 4,000 base pairs amplicon of a near full-length 16S-ITS-23S gene) was first used by Martijn et al. (2019) and showed potential to cover anaerobic microbes including bacteria and methanogens, which are most important for wastewater treatment processes. However, this study found that the primer pair requires further optimisation to cover a greater diversity of wastewater microbes.

We compared the taxonomic profiles derived from the whole operon primers in this study with more common primers that were used on the same sludge as part of a separate study. Microbial profiling with the V3-V4 primer pair (a validated primer, targeting a short region within the 16S rRNA gene) agreed well with the microbial profiles of metagenomes of the same sludge, showing that the V3-V4 primer provided a realistic representation of microbial diversity (albeit only to genus-level) (Fig. 7 and 8). Hence, taking the V3-V4 data and metagenome as a baseline, it was apparent that the whole operon primer pair used in this study did not provide sufficient coverage of process-relevant microbes in anaerobic sludge (Fig. 7 and 8).

The results showed that the use of primers A519F/ U2428R resulted in a strong bias for two types of microbes, the *Anaerolineaceae* bacteria and *Methanobacterium* methanogens (Fig. 7 and 8). Some key bacteria and archaea were underrepresented, including *Methanothrix* and *Methanosarcina*, two important methanogens and marker microbes for healthy digester operation (Fig. 8). It therefore appears that using the 16S-ITS-23S primer pair did not result in a true representation of microbial abundances.

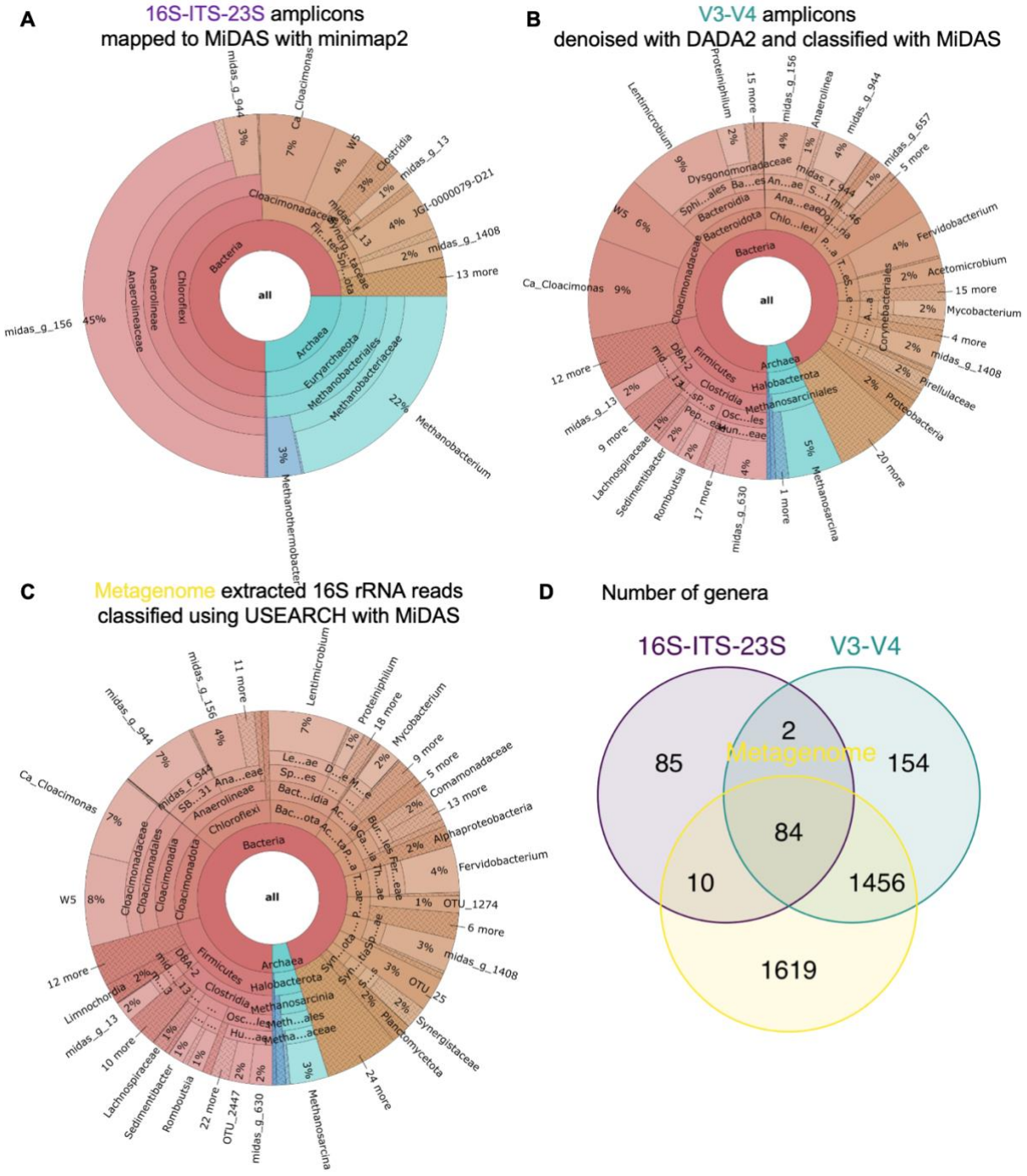

Figure 7. Krona plot comparison of relative abundances of (**A**) 16S-ITS-23S amplicons, with (**B**) V3-V4 amplicons and (**C**) metagenome-extracted full-length 16S reads. All sequences were classified with the MiDAS 5.3 database and abundances normalised (rarefied to lowest sample sum with replacement). DNA for A was extracted and sequenced using the same sludge samples as for B and C. More details about B and C are available in Krohn et al. (2025, under review). Some sequences were given a 'midas' placeholder if identities were unknown. (**D**) The shared number of genera between A, B and C are shown.

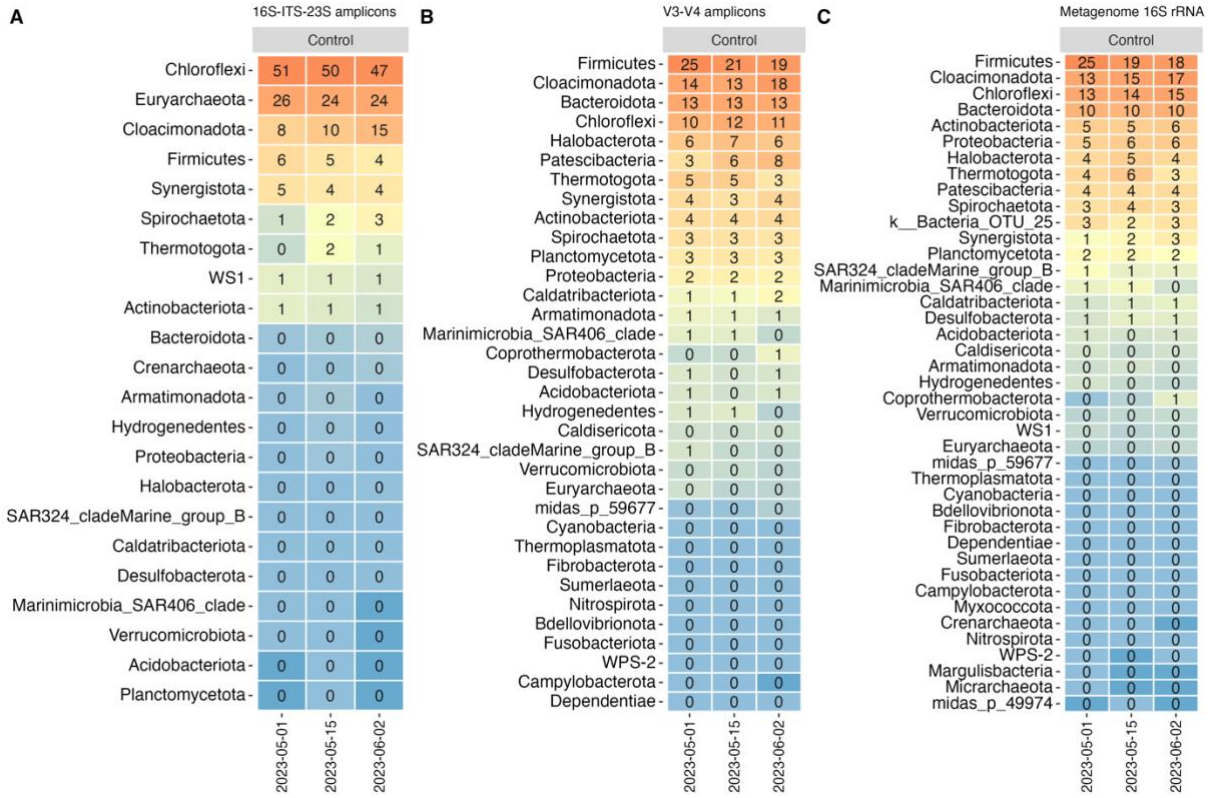

Figure 8. Comparison of relative abundances (%) of (**A**) 16S-ITS-23S amplicons, with (**B**) V3-V4 amplicons and (**C**) metagenome-extracted full-length 16S reads, aggregated to phylum level. All sequences were classified with the MiDAS 5.3 database and abundances normalised (rarefied to lowest sample sum with replacement). DNA for A was extracted and sequenced using the same sludge samples as for B and C. 16S-ITS-23S amplicons were mapped to database sequences with minimap2. V3-V4 amplicons were denoised with DADA2 prior to taxonomic classification. Metagenome full-length 16S reads were clustered and classified with USEARCH11. More details about B and C are available in Krohn et al. (2025, under review).

4. Recommendations for developing a digester health test

This study showed that whole operon amplicon sequencing has the potential to significantly improve diagnostics and the monitoring of biological processes and pathogen risks. This approach has the potential to detect the presence of hundreds of species and even strains and indicate their relative changes over time during treatment processes. It can be developed into a routine diagnostic test and cover multiple use-cases for process optimisation.

We proposed this sequencing approach for the modelling of temporal changes of species to predict process outputs such as methane concentrations. However, before whole operon sequencing can be utilised, we recommend to first optimise primers. As shown in the results, the primer did not cover all the process-relevant microbes, hence would make it less predictive for the process, compared to existing primers (which cover more types of microbes but not at species-level).

The primer used for this study covered only a subset of known wastewater species, which is nonetheless a valuable starting point for testing several variations of the current primer sequences. The development

of alternative primers may also be necessary if optimisation of existing primers is not successful. First, primers can be tested *in silico* and then tested on waste-activated, anaerobic sludge and on standard reference communities to benchmark their performance.

Databases may also need validation. Whole operon sequencing is a novel approach, and whole operon reference sequences are not commonly used yet, especially in engineered wastewater environments. Hence, it is recommended to further validate the sequences of any species identified with the GROND database by aligning them to existing genomes and the MiDAS reference sequences to rigorously verify their identity. Culturing of novel sludge isolates to sequence and annotate their genome may further help in this endeavor.

One challenge of whole operon sequencing is the variability of the ITS region (Brewer et al., 2020), with some microbial lineages having extremely long (>1,500 bp) or 'unlinked' ITS genes, potentially resulting in their exclusion from the data. We recommend to systematically assess existing wastewater genomes for length of ITS region and understand the extend of this phenomenon for wastewater microbes and, if needed, develop strategies to manage this in the ongoing primer design.

The ITS region may also be a valuable marker for difficult-to-discriminate species, including Cyanobacterial species (Boyer et al., 2001). Hence, a bioinformatic approach could be developed that extract each of the loci (16S, ITS, 23S) from the amplicon reads to evaluate them separately and come to a consensus solution for taxonomic classification.

As the above work requires significant dedication it is likely best achieved through a PhD project.

To progress the digester health test, we propose the use of existing V3-V4 or V4 primer pairs. Although, they do not provide species-level detection, they cover a much greater diversity and thus are likely able to discriminate some of the process outputs. For example, preliminary assessments of RMIT laboratory-scale digesters with V3-V4 primers have shown that temporal changes of methanogens agreed well with changes in methane concentrations.

References

Adhikary, S., Nicklas, W., Bisgaard, M., Boot, R., Kuhnert, P., Waberschek, T., Aalbæk, B., Korczak, B., Christensen, H., 2017. Rodentibacter gen. nov. including Rodentibacter pneumotropicus comb. nov., Rodentibacter heylii sp. nov., Rodentibacter myodis sp. nov., Rodentibacter ratti sp. nov., Rodentibacter heidelbergensis sp. nov., Rodentibacter trehalosifermentans sp. nov., Rode. Int. J. Syst. Evol. Microbiol. 67, 1793–1806.

Bayley, H., 2015. Nanopore sequencing: From imagination to reality. Clin. Chem. 61, 25–31. https://doi.org/10.1373/clinchem.2014.223016

Boyer, S.L., Flechtner, V.R., Johansen, J.R., 2001. Is the 16S-23S rRNA internal transcribed spacer region a good tool for use in molecular systematics and population genetics? A case study in cyanobacteria. Mol. Biol. Evol. 18, 1057–1069. https://doi.org/10.1093/oxfordjournals.molbev.a003877

Brewer, T.E., Albertsen, M., Edwards, A., Kirkegaard, R.H., Rocha, E.P.C., Fierer, N., 2020. Unlinked rRNA genes are widespread among bacteria and archaea. ISME J. 14, 597–608. https://doi.org/10.1038/s41396-019-0552-3

Brune, A., Song, Y., Oren, A., Paster, B.J., 2022. A new family for 'termite gut treponemes': description of Breznakiellaceae fam. nov., Gracilinema caldarium gen. nov., comb. nov., Leadbettera azotonutricia gen. nov., comb. nov., Helmutkoenigia isoptericolens gen. nov., comb. nov., and Zuelzera stenostre. Int. J. Syst. Evol. Microbiol. 72, 5439. https://doi.org/10.1099/ijsem.0.005439

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583.

Chang, S.C., Kao, M.R., Saldivar, R.K., Díaz-Moreno, S.M., Xing, X., Furlanetto, V., Yayo, J., Divne, C., Vilaplana, F., Abbott, D.W., Hsieh, Y.S.Y., 2023. The Gram-positive bacterium Romboutsia ilealis harbors a polysaccharide synthase that can produce (1,3;1,4)-β-d-glucans. Nat. Commun. 14, 4526. https://doi.org/10.1038/s41467-023-40214-z

Curry, K.D., Wang, Q., Nute, M.G., Tyshaieva, A., Reeves, E., Soriano, S., Wu, Q., Graeber, E., Finzer, P., Mendling, W., Savidge, T., Villapol, S., Dilthey, A., Treangen, T.J., 2022. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat. Methods 19, 845–853. https://doi.org/10.1038/s41592-022-01520-4

De Coster, W., Rademakers, R., 2023. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics 39, btad311. https://doi.org/10.1093/bioinformatics/btad311

Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524. https://doi.org/10.1038/nbt.3423

Dyksma, S., Gallert, C., 2019. Candidatus Syntrophosphaera thermopropionivorans: a novel player in syntrophic propionate oxidation during anaerobic digestion. Environ. Microbiol. Rep. 11, 558–570.

Krohn, C., Jansriphibul, K., Dias, D.A., Rees, C.A., Akker, B. van den, Boer, J.C., Plebanski, M., Surapaneni, A., O'Carroll, D., Richard, S., Batstone, D.J., Ball, A.S., 2024. Dead in the water – Role of relic DNA and primer choice for targeted sequencing surveys of anaerobic sewage sludge intended for biological monitoring. Water Res. 253, 121354. https://doi.org/10.1016/j.watres.2024.121354

Li, H., 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., Steinegger, M., 2022. Metagenome analysis using the Kraken software suite. Nat. Protoc. 17, 2815–2839. https://doi.org/10.1038/s41596-022-00738-y

Martijn, J., Lind, A.E., Schön, M.E., Spiertz, I., Juzokaite, L., Bunikis, I., Pettersson, O. V., Ettema, T.J.G., 2019. Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. Environ. Microbiol. 21, 2485–2498. https://doi.org/10.1111/1462-2920.14636

Mavromatis, K., Stackebrandt, E., Held, B., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.F., Tapia, R., Goodwin, L.A., Pitluck, S., Liolios, K., Pagani, I., Ivanova, N., Mikhailova, N., Huntemann, M., Pati, A., Chen, A.,

Palaniappan, K., Land, M., Rohde, M., Spring, S., Göker, M., Woyke, T., Detter, J.C., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Klenk, H.P., Kyrpides, N.C., 2013. Complete genome sequence of the moderate thermophile Anaerobaculum mobile type strain (NGAT). Stand. Genomic Sci. 8, 47–57. https://doi.org/10.4056/sigs.3547050

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P.M., Stevens, M.H.H., & Wagner, H., 2008. The vegan package. Community Ecol. Packag.

Pelletier, E., Kreimeyer, A., Bocs, S., Rouy, Z., Gyapay, G., Chouari, R., Rivière, D., Ganesan, A., Daegelen, P., Sghir, A., Cohen, G.N., Médigue, C., Weissenbach, J., Le Paslier, D., 2008. "Candidatus Cloacamonas acidaminovorans": Genome sequence reconstruction provides a first glimpse of a new bacterial division. J. Bacteriol. 190, 2572–2579. https://doi.org/10.1128/JB.01248-07

Pohlschroeder, M., Leschine, S.B., Canale-Parola, E., 1994. Spirochaeta caldaria sp. nov., a thermophilic bacterium that enhances cellulose degradation by Clostridium thermocellum. Arch. Microbiol. 161, 17–24. https://doi.org/10.1007/BF00248889

Rang, F.J., Kloosterman, W.P., de Ridder, J., 2018. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19, 90. https://doi.org/10.1186/s13059-018-1462-9

Rodríguez-Pérez, H., Ciuffreda, L., Flores, C., 2021. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. Bioinformatics 37, 1600–1601. https://doi.org/10.1093/bioinformatics/btaa900

Seol, D., Lim, J.S., Sung, S., Lee, Y.H., Jeong, M., Cho, S., Kwak, W., Kim, H., 2022. Microbial identification using rRNA operon region: Database and tool for metataxonomics with long-read sequence. Microbiol. Spectr. 10, e02017-21. https://doi.org/10.1128/spectrum.02017-21

Walsh, C.J., Srinivas, M., Stinear, T.P., van Sinderen, D., Cotter, P.D., Kenny, J.G., 2024. GROND: a quality-checked and publicly available database of full-length 16S-ITS-23S rRNA operon sequences. Microb. genomics 10. https://doi.org/10.1099/mgen.0.001255

Wang, Yunhao, Zhao, Y., Bollas, A., Wang, Yuru, Au, K.F., 2021. Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol. 39, 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20, 257. https://doi.org/10.1186/s13059-019-1891-0

Wright, R.J., Comeau, A.M., Langille, M.G.I., 2023. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microb. Genomics 9, 949. https://doi.org/10.1099/mgen.0.000949