Database, 2022, 1–9
DOI: https://doi.org/10.1093/database/baac041
Original article

AcetoBase Version 2: a database update and re-analysis of formyltetrahydrofolate synthetase amplicon sequencing data from anaerobic digesters

Abhijeet Singh* and Anna Schnürer*

Department of Molecular Sciences, BioCenter, Anaerobic Microbiology and Biotechnology Group, Swedish University of Agricultural Sciences, Almas Allé 5, Uppsala SE-750 07, Sweden

*Correspondence may also be addressed to Anna Schnürer. Tel: +46 18671000; Fax: +46 18672000; Email: anna.schnurer@slu.se and Abhijeet Singh. Tel: +46 18671000; Fax: +46 18672000; Email: abhijeet.singh@slu.se

Citation details: Singh, A. and Schnürer, A. AcetoBase Version 2: a database update and re-analysis of formyltetrahydrofolate synthetase amplicon sequencing data from anaerobic digesters. *Database* (2022) Vol. 2022: article ID baac041; DOI: https://doi.org/10.1093/database/baac041

Abstract

AcetoBase is a public repository and database of formyltetrahydrofolate synthetase (FTHFS) sequences. It is the first systematic collection of bacterial FTHFS nucleotide and protein sequences from genomes and metagenome-assembled genomes and of sequences generated by clone library sequencing. At its publication in 2019, AcetoBase (Version 1) was also the first database to establish connections between the FTHFS gene, the Wood–Ljungdahl pathway and 16S ribosomal RNA genes. Since the publication of AcetoBase, there have been significant improvements in the taxonomy of many bacterial lineages and accessibility/availability of public genomics and metagenomics data. The update to the AcetoBase reference database described here (Version 2) provides new sequence data and taxonomy, along with improvements in web functionality and user interface. The evaluation of this latest update by re-analysis of publicly accessible FTHFS amplicon sequencing data previously analysed with AcetoBase Version 1 revealed significant improvements in the taxonomic assignment of FTHFS sequences.

Database URL: https://acetobase.molbio.slu.se

Introduction

Formyltetrahydrofolate synthetase (FTHFS) is a key marker gene of the Wood–Ljungdahl pathway (WLP) of acetogenesis (1, 2). In the enzymatic process of acetogenesis, FTHFS facilitates ATP-dependent conversion of formate into formyltetrahydrofolate in the methyl branch of the WLP. Although the FTHFS gene is also present in other bacteria, e.g. syntrophic acid-oxidizing bacteria, sulphate-reducing bacteria and methanogens, only acetogens use it in the true sense of acetogenesis (1, 3–5). For extensive culture-independent investigations into the ecology of acetogenic communities in different environments, modern molecular methods are of immense importance (1). In this context, FTHFS has been used for decades as a molecular marker for the identification of potential acetogenic bacterial communities (1, 6–8). As discussed in several earlier studies, acetogens are phylogenetically diverse, metabolically very agile and important in many environments (3, 5, 9, 10). Recently, we developed a high-throughput sequencing and analysis methodology of FTHFS gene sequencing and compared it to other frequently used methods for microbial community analysis in biogas environments (4, 11). The application of our analysis strategy extensively revealed the structural composition and temporal dynamics of the potential acetogenic communities in different anaerobic digesters using different feed substrates (11, 12).

Accurate analysis of high-throughput amplicon sequencing data relies strongly on the quality and taxonomic accuracy of the reference database. For reliable identification and annotation of high-throughput amplicon sequencing data, we have created and published AcetoBase, the first curated database and repository of FTHFS sequences (3). Since its initial publication, there has been a significant increase in the number of genomic and metagenome-assembled genome (MAG) data sets containing FTHFS sequences deposited in public databases. In addition, following recent technological improvements in genomics and bioinformatics, the taxonomy of many bacterial lineages has been re-defined. Thus, for correct identification and annotation of FTHFS sequences, it is relevant and necessary to update AcetoBase with recent and updated information. In this paper, we describe recent changes and updates made to AcetoBase and present results from re-analysis of the FTHFS sequence data generated in our previous studies.

Received 20 January 2022; Revised 3 May 2022; Accepted 04 May 2022

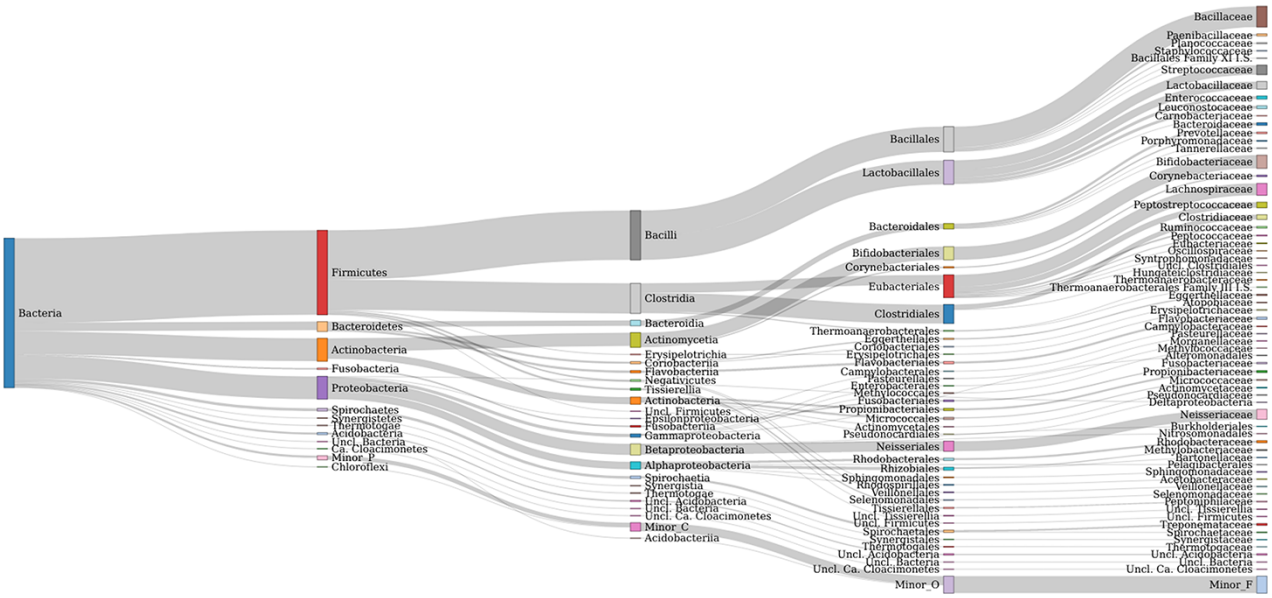

Figure 1. Taxonomic coverage up to the family level of FTHFS sequences present in the AcetoBase reference protein data set. Candidates with <50 sequences are merged in minor taxa.

New developments

Database content

AcetoBase Version 2 contains ∼26 000 protein and ∼20 000 nucleotide sequences retrieved from public repositories. These sequences belong to 8439 distinct taxonomic identifiers. In comparison, Version 1 of AcetoBase contained ∼18 000 protein and ∼13 000 nucleotide sequences, belonging to 7 928 taxonomic identifiers. Thus, the new update has increased the AcetoBase sequence (protein and nucleotide) and taxonomic identifier diversity by ∼7 300, 7000 and 500, respectively. The major phyla associated with AcetoBase Version 2 sequences are Firmicutes, Actinobacteria, Proteobacteria and Bacteroidetes (Figure 1). An interactive version of Figure 1 is available at https://acetobase.molbio.slu.se/home/sankey. Most sequences (nucleotide and protein) in AcetoBase Version 1 were from isolated or characterized bacteria. However, in addition to sequences from bacterial isolates, in AcetoBase Version 2 FTHFS sequences from MAGs are also included. The FTHFS nucleotide and translated protein sequences belonging to MAGs do not have valid accession numbers; thus, they could not be associated with any valid identifier in National Center for Biotechnology Information (NCBI) databases. Therefore, these sequences are included as user nucleotide (UN_0000025797-UN_0000029649) and user protein (UP_0000000001-UP_0000003853) sequences in respective AcetoBase database tables while preserving complete information on the sequence origin. The FTHFS nucleotide and translated protein sequences from MAGs are included as published and new sequences, respectively, with AcetoBase as sequence owner/user identity.

Approximately 3100 clone sequences (47 new) are present in the AcetoBase clone data set. In AcetoBase Version 1, taxonomy prediction for the clone sequences was performed using the SINTAX algorithm (13). However, in AcetoBase Version 2, the taxonomy for the complete clone data set is predicted with the acetotax program in the AcetoScan pipeline (4) and using the updated AcetoBase reference protein data set,

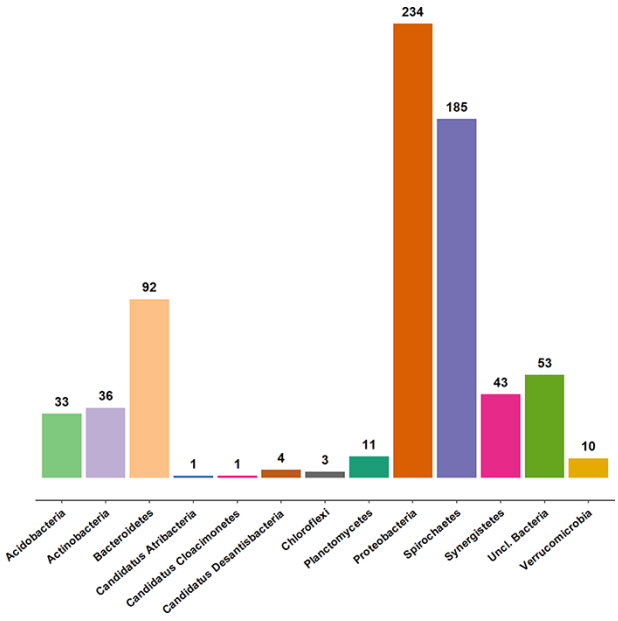

Figure 2. Taxonomic coverage of the clone sequence data set of AcetoBase at the phylum level, excluding Firmicutes ($n = 2355$ sequences). The values shown represent the sequence count for the respective phylum present in the AcetoBase clone data set.

which is accessible at https://acetobase.molbio.slu.se/download/ref/1. Acetotax is an unsupervised sequence annotation program that filters out non-FTHFS sequences, performs best open reading frame (ORF) analysis and annotates taxonomy to FTHFS sequences based on the latest version of AcetoBase. The new taxonomic prediction of clone sequences (best ORF) indicates that most clones generated for FTHFS sequences belong to three major phyla, i.e. Firmicutes (2355/3061 sequences, 77%), Proteobacteria and Spirochaetes. The taxonomic coverage of clone sequences present in AcetoBase is presented in Figure 2. For ease of visualization, sequences

Figure 3. Screenshot of taxonomic annotation and placement of clone sequences associated with the study by Parameswaran *et al.* (14) in the AcetoBase clone phylogenetic tree available at https://acetobase.molbio.slu.se/phylo/clone.

($n = 2355$) associated with the Firmicutes are not included in the diagram.

Taxonomy prediction for the clone sequences helped in correctly associating the clone sequences to a bacterial species. For instance, it revealed that, among the sequences generated in a study by Parameswaran *et al.* (14), none of the sequences submitted as uncultured *Alkaliphilus* sp. clone (10 clone sequences) belong to *Alkaliphilus*. In fact, 46/47 clone sequences available from that study were found to be more closely related to *Acetobacterium* spp. (>94% blastx similarity) (Figure 3). These new clone sequences, with putative taxonomy and species-level percentage identity, are available in the AcetoBase clone database under accession numbers CN_0000003015–CN_0000003061. The phylogenetic trees for the reference protein, nucleotide and clone data sets have also been reconstructed according to the updated database content, and information about phylogenetic tree construction is now provided at the web interface.

AcetoBase reference data sets

AcetoBase reference sequence databases (protein and nucleotide) were constructed from sequences present in reference tables (NP and NN) and new sequences from MAGs as described earlier (UP and UN). For the reference protein database, the NP and UP table sequences were merged, and then duplicate sequences (100% sequence similarity) were removed with the program DupRemover (15) and sequences with <300 amino acids were removed with the program FilterByLength (16). For the reference nucleotide database, the NN and UN table sequences were also merged, and then redundant sequences were removed with the program DupRemover and sequences with <1000 bases were removed with the program FilterByLength. The reference protein and nucleotide data sets in AcetoBase Version 2 contain 12 970 and 10 266 sequences and are available under the names acetobase_ref_Prot_latest.tar.gz and acetobase_ref_Nucl_latest.tar.gz, respectively. These reference databases, along with their previous versions, are available as archive databases and can be accessed at the download page https://acetobase.molbio.slu.se/download.

The latest AcetoBase data sets (protein and nucleotide) were used for the construction of phylogenetic trees by replacing invalid sequence residues with Xs in protein and Ns in nucleotide sequences using the program clean_AMINOACID_fasta (17) and clean_DNA_fasta (18), respectively. Sequences were clustered at 80% sequence similarity using VSEARCH (19) and aligned with FAMSA (20). Phylogenetic trees were then constructed with FastTree2 (21). For the construction of a clone phylogenetic tree, the taxonomically annotated clone sequence data set (described earlier), also containing both AcetoBase and NCBI accession numbers, was filtered and duplicate sequences were removed with DupRemover, remaining sequences were aligned with FAMSA and trees were constructed with FastTree2. The phylogenetic trees created (phylogenetic tree tip count: protein = 2361, nucleotide = 3941 and clone = 3048) are available on the phylogeny page. For reproducibility, the commands used for the preparation of respective trees are accompanied by interactive phylogenetic trees at the AcetoBase phylogeny web interface.

RibocetoBase data set

In AcetoBase Version 1, FTHFS sequences were connected to 16S ribosomal RNA (rRNA) genes for the respective bacteria, but these sequences were not made available as a reference data set for taxonomic annotations. In AcetoBase Version 2, we have included RibocetoBase, a 16S rRNA gene sequence data set of the FTHFS-harbouring bacterial community that contains ∼9169 sequences, which can

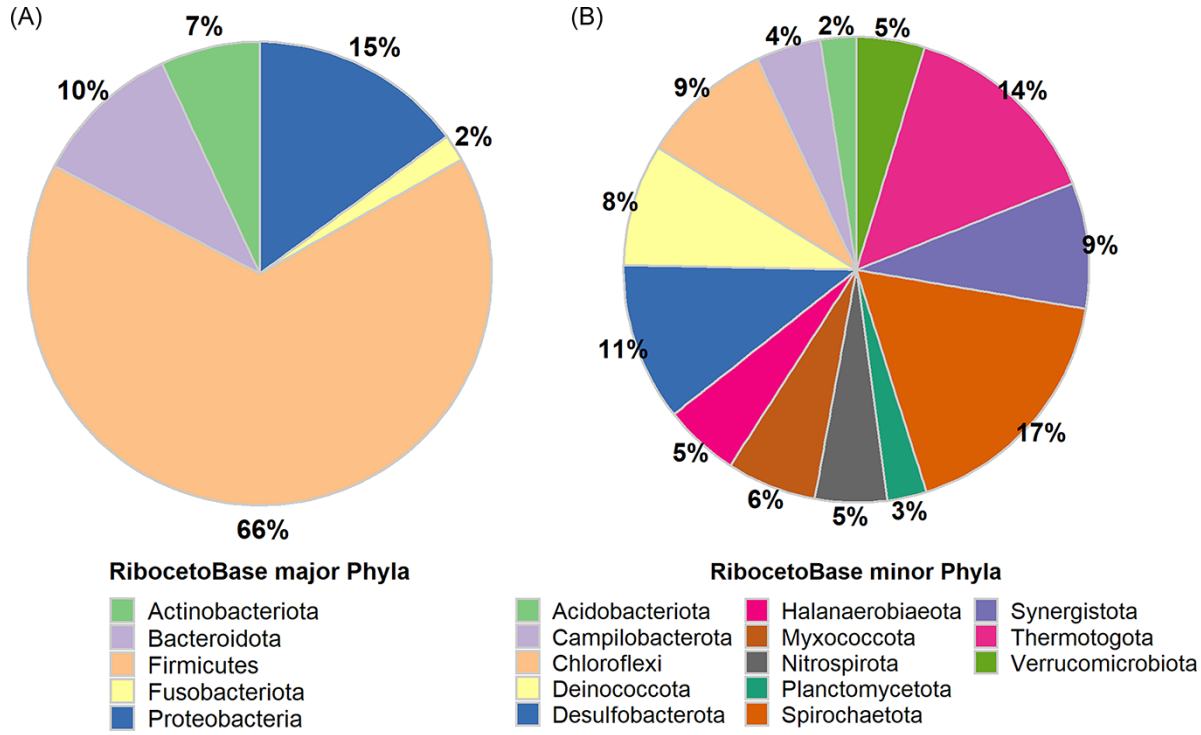

Figure 4. Taxonomic distribution in the RibocetoBase data set of FTHFS-harbouring communities present in AcetoBase at the phylum level. (A) Major phyla with >100 sequences and (B) minor phyla with sequence count 10–100.

be readily downloaded and used for taxonomic annotation of the FTHFS-containing community from 16S rRNA gene sequences. Full details of RibocetoBase and its development can be found elsewhere (11). The taxonomic coverage of the five major phyla (count >100 sequences, 8733/9169) and 13 minor phyla (count 10–100 sequences, 401/9169) are presented in Figure 4A and B, respectively. Around 20 phyla with <10 sequences (35/9169) are ignored in the visualization in Figure 4. RibocetoBase taxonomic lineages are according to the Genome Taxonomy Database (22). The RibocetoBase data set can be accessed at https://acetobase.molbio.slu.se/download/ref/3.

Enhanced web interface and functionalities

FTHFS has long been used as a marker gene for acetogens, and several acetogens have been isolated and characterized over the years. However, a detailed inventory of the original articles and bacterial micrographs in which these acetogens were first isolated, characterized and described was lacking. In AcetoBase Version 2, we have included a bacterial micrograph collection of well-characterized acetogens ($n = 52$) whose FTHFS is available in AcetoBase and of acetogens proposed in the literature ($n = 64$), which are either not taxonomically validated or lack genome sequences. For these acetogens, links are provided in AcetoBase Version 2 to the original articles describing the isolation and description of the acetogen. A checkmark represents the original article associated with the respective acetogen, and in cases, where the original article describing the isolation and characterization of an acetogen does not have a micrograph, images from other publicly available articles have been retrieved. We have also linked the acetogen to the NCBI taxonomy database, which can be helpful in (i) getting complete information on different strains of the respective species, (ii) compiling taxonomic names and synonyms and their state of validation and (iii) linking out to other NCBI databases via the NCBI taxonomy database, for extra information. The AcetoBase Version 2 information resource for acetogens is the most descriptive and elaborate collection of acetogens to date and is accessible at https://acetobase.molbio.slu.se/organism/acetogen.

Although FTHFS is commonly used as a marker for acetogens, as mentioned earlier it is present in some syntrophic acid-degrading and sulphate-reducing bacteria (1, 3, 4). In several previous studies on FTHFS gene amplicon-based community profiling (2, 7, 8, 11, 23, 24), syntrophic acid-degrading bacteria and sulphate-reducing bacteria have been reported together with acetogenic bacteria. These groups of bacteria use the FTHFS gene in e.g., biosynthesis of folate or short-chain fatty-acid oxidation but not in reductive acetogenesis, unlike the acetogens. Thus, considering the prospects and potential for FTHFS gene profiling in different environments, in AcetoBase Version 2 we have included information (similarly to that described earlier for acetogens) on FTHFS-harbouring syntrophic acid-degrading and sulphate-reducing bacteria. It is worth noting that not all syntrophic acid-degrading bacteria and sulphate-reducing bacteria possess the FTHFS gene in their genome and that AcetoBase Version 2 only includes bacterial species harbouring the FTHFS gene. We have also grouped these bacteria according to the functional categories in which they have been described, namely syntrophic acetate-oxidizing bacteria, syntrophic propionate-oxidizing bacteria, syntrophic butyrate-oxidizing bacteria, syntrophic fatty-acid-oxidizing bacteria, sulphate-reducing bacteria and syntrophic benzene-oxidizing

Nucleotide Sequences (NN)

AcetoBase ID	Genome ID	Protein ID	Lineage	Species	Strain
NN_0000012586	NZ_AXAC01000015.1	WP_026394331.1	p: Firmicutes |_c: Clostridia |_o: Eubacteriales |_f: Eubacteriaceae |_g: Acetobacterium	Acetobacterium dehalogenans	None

Protein Sequences (NP)

AcetoBase ID	Protein ID	Gene	Lineage	Species	Strain
NP_0000014201	WP_026394331.1	FTHFS_1	p: Firmicutes |_c: Clostridia |_o: Eubacteriales |_f: Eubacteriaceae |_g: Acetobacterium	Acetobacterium dehalogenans	None
NP_0000017748	WP_026394331.1	FTHFS_1	p: Firmicutes |_c: Clostridia |_o: Eubacteriales |_f: Eubacteriaceae |_g: Acetobacterium	Acetobacterium dehalogenans	None

Clone Sequences (CN)

AcetoBase ID	Clone ID	Protein ID	Source	Putative taxonomy	Definition
CN_0000002287	KF928182.1	AHC28486.1	production water from oil reservoir	p: Firmicutes |_c: Clostridia |_o: Eubacteriales |_f: Eubacteriaceae |_g: Acetobacterium |_s: Acetobacterium dehalogenans (97.436)	Uncultured Clostridium sp. clone FTHFS S5 5 formyltetrahydrofolate synthetase FTHFS gene partial cds
CN_0000003061	FJ848965.1	ACP43437.1	microbial fuel cell using wastewater inoculum from Mesa Northwest Wastewater Reclamation Plant	p: Firmicutes |_c: Clostridia |_o: Eubacteriales |_f: Eubacteriaceae |_g: Acetobacterium |_s: Acetobacterium dehalogenans (96.129)	Uncultured Clostridium sp. clone D12 formyltetrahydrofolate synthetase gene partial cds
CN_0000003024	FJ848928.1	ACP43400.1	microbial fuel cell using wastewater inoculum from Mesa Northwest Wastewater Reclamation Plant	p: Firmicutes |_c: Clostridia |_o: Eubacteriales |_f: Eubacteriaceae |_g: Acetobacterium |_s: Acetobacterium dehalogenans (94.737)	Uncultured Alkaliphilus sp. clone F3 formyltetrahydrofolate synthetase gene partial cds

Figure 5. Screenshot of search results in AcetoBase Version 2 showing taxonomic lineages of the query against the nucleotide and protein data sets. The search results for the clone data set show the isolation source of the clone, its description/definition and the putative taxonomy predicted with the acetotax program.

bacteria. This systematic grouping and collection of syntrophic bacteria can be accessed at https://acetobase.molbio.slu.se/organism/syntroph.

Search functionality and results

The new version of AcetoBase now supports searches with a query from any taxonomy level (phylum-strain) and regular expressions (matching word patterns) in the search bar. The search results are presented with a link out functionality to the respective NCBI database for genome and protein identifiers. The complete lineage for the search results is displayed with link out to the NCBI taxonomy database from the species name (Figure 5). Wherever possible, strain-level information is provided for bacterial species and MAGs. With respect to the clone data set, the search function now also supports queries with taxa names, regular expressions and searches with isolation source variables, e.g. anaerobic digester, gut, roots, colon and microbial fuel cell (Figure 5). The search results for the clone data set now also present the putative taxonomy predicted for the clone sequences, as mentioned earlier. The species-level percentage identity is available with the lineage for each clone identifier (Figure 5).

Re-analysis of FTHFS high-throughput sequencing data

AcetoBase Version 1 was published in 2019 and was used for high-throughput sequencing data analysis in our previous studies (4, 11, 24). To evaluate the impact of the update to AcetoBase on the taxonomic annotation of the FTHFS operational taxonomic units (OTUs) generated by the AcetoScan pipeline, we re-analysed the data from our three previous studies using AcetoBase Version 2. The methodology used in the re-analysis is described in the following section.

Methodology

Re-analysis was carried out on data associated with the articles by Singh *et al.* (4) (data set 1), Singh *et al.* (11) (data set 2) and Singh *et al.* (24) (data set 3). These data sets were re-analysed separately with the AcetoScan pipeline, using the parameters described in the respective paper and the AcetoBase Version 2 protein data set as the reference database. The results from re-analysis of the respective data set were compared with results from previous analyses, based on differences and variations in the taxonomic annotation. For ease of understanding, in the following text, the expression

'previous study/analysis' refers to results published for the respective data set/paper, and 're-analysis' refers to results of the re-analysis of the respective data set using AcetoBase Version 2 in the present study. Specific details of the re-analysis and major community insights gained are presented in the following sub-sections.

Re-analysis of data set 1

In re-analysis of data set 1, in addition to results from the previous study (4), we also included a new set of FTHFS sequence data for reactor GR1 that was not presented in that study, which only used data from reactor GR2 (Supplementary Figure S1A–S1F). The analysis in the original article was performed separately for the forward and reverse reads. Our re-analysis for both reactors GR1 and GR2 was conducted by merging the forward and reverse reads of each respective sample. The analysis was performed with a 100% clustering threshold and using AcetoScan analysis parameters -m 300, -n 120, -q 20, -c 5 and -t 1.0. The previous study described 935 and 662 OTUs obtained at a 100% clustering threshold from forward and reverse reads in reactor GR2, respectively. The re-analysis of data set 1 and the new data for reactor GR1 resulted in 1025 OTUs at a 100% clustering threshold when using merged data from forward and reverse reads. The re-analysis results indicated a higher percentage of annotated sequences compared with the previous analysis and with no unclassified OTUs at the genus level (Supplementary Figure S1E). Genera related to OTUs that were not classified in the re-analysis, but were classified in the previous analysis, were *Butyrivibrio*, *Caldisalinibacter*, *Eggerthella*, *Hungateiclostridium*, *Lagierella*, *Maledivibacter* and *Phocea*. Taxa classified in the re-analysis but not classified in the previous study included, e.g., *Peptoniphilus*, *Ruminiclostridium*, *Sedimentibacter*, *Thermoanaerobacter* and *Urinicoccus*. The genus *Senegalimassilia* was only observed in the new additional data from reactor GR1 and not in the data from reactor GR2 used in the previous study (Supplementary Figure S1E).

Re-analysis of data set 2

Re-analysis of data set 2 was done for forward reads as described in the original article (11), with the AcetoScan analysis parameters -m 300, -n 150, -q 21, -c 10, -r 1, -e 1e-30 and -t 1.0. The number of OTUs ($n = 387$) generated in the re-analysis was similar to that generated in the previous study ($n = 391$). The re-analysis results (Supplementary Figure S2A–S2F) for data set 2 at the phylum level showed a significant reduction in the annotation of OTUs classified previously as Actinobacteria and, in contrast to the previous analysis, no annotations were obtained for Spirochaetes at relative abundance (RA) >1% (Supplementary Figure S2A). Compared with the previous study, the fraction of OTUs classified as Firmicutes increased on re-analysis. At the genus level, genera that were not classified in the re-analysis were *Caloranaerobacter*, *Methylocystis*, *Oscillibacter*, *Phocea*, *Tissierella*, *Treponema* and *Varibaculum* (Supplementary Figure S2E). Clear and significant differences in the taxonomic annotations were noted for the OTUs previously annotated as *Varibaculum*, which in the re-analysis were annotated as unclassified *Urinicoccus*. The total number of genera (RA >1%) annotated in the re-analysis was 17, compared with 21 in the previous study. Apart from the changes at the genus level, the most significant change in the re-analysis was for species belonging to the phylum Cloacimonetes (Supplementary Figure S2F). In the re-analysis, only one species from this phylum was classified, i.e. Candidatus Cloacimonetes bacterium (MAG ID: AS05jafATM_99) originating from anaerobic digesters (25). In comparison, the previous study annotated two species, Cloacimonetes bacterium HGW-Cloacimonetes-1 and Cloacimonetes bacterium HGW-Cloacimonetes-2, which originated from a metagenomics project analysing subsurface sediments (26).

Re-analysis of data set 3

Data set 3 was generated in a study by Singh *et al.* (24) involving high-throughput microbiological surveillance of six different biogas plants (one plug flow and five parallel continuous stirred-tank reactors plants) with a total of 11 biogas reactors. The AcetoScan analysis parameters used in the re-analysis were the same as those used in the previous study (-m 300, -n 150, -q 20, -c 5, -e 1e-30 and -t 1.0). The re-analysis resulted in 1899 OTUs, compared with 1901 OTUs in the previous study (Supplementary Figure S3A–S3F). At the phylum level, the re-analysis results revealed similar community diversity in all reactors as previously reported, except for reactor C6. The re-analysis results showed an increase in the percentage of OTUs annotated as Firmicutes, with a corresponding decrease in OTUs for the phylum Actinobacteria. A minor reduction was observed for OTUs previously belonging to Synergistetes, which were not classified as Synergistetes in the re-analysis (Supplementary Figure S3A). Due to the high numbers of genera in both the previous study and the re-analysis, the results obtained for each genus and species are not included here (see Supplementary Figures S3E and S3F for comparison). However, an overview of the annotation results indicated that annotation accuracy improved significantly in the re-analysis, with a smaller number of resulting taxa and higher percentage identity to the reference database than in the previous study.

Discussion

The FTHFS sequence data generated in previous studies were re-analysed here using the updated Version 2 of AcetoBase, and differences in taxonomic annotation were compared. Careful interpretation of the re-analysis results and cross-comparison to the previous studies (4, 11, 24) supported and justified the improvements in taxonomic annotations resulting from the updates to AcetoBase. The increased taxonomic diversity in the reference database was shown to provide an opportunity for taxonomic annotation of OTUs with higher accuracy (greater percentage identity and lower *e*-value) in the best-hit strategy used by AcetoScan. However, since FTHFS is not a taxonomic marker and due to technical limitations in FTHFS amplicons sequencing, the community diversity evaluated with the FTHFS gene can have similarities, but also differences, when compared to the 16S rRNA gene and whole-genome metagenome-based community profiles. Nevertheless, a comparative analysis of FTHFS and 16S rRNA gene amplicons by Singh *et al.* (11) demonstrated that the microbial community profiled with FTHFS showed high similarity to the structure and dynamics revealed by community profiling using the 16S rRNA gene. The recent changes in the taxonomic lineages of bacteria also contributed to differences

between the re-analysis results and those in the previous studies.

Scope of acetobase

Acetogenic bacterial communities are metabolically diverse and are found in hugely varied environments such as anaerobic digesters, animal gut, rumen, human gut, marine sediments, forest soil, peatlands, permafrost and microbial electrochemical systems (10, 11, 14, 24, 27–34). In these environments, they are not only involved in the carbon cycling via acetate but are also extensively involved in the metabolism of ethanol, methanol, formate, butyrate, lactate, vanillate, chlorinated compounds, cellulose, mono-tetra saccharides and methylated amines (2, 35–43).

AcetoBase was initially developed for use in the analysis of the acetogenic community in anaerobic digestor/biogas environments. However, AcetoBase contains large numbers of FTHFS-harbouring bacterial species that are also present in other natural environments. As discussed earlier, FTHFS primers can target syntrophic acid-oxidizing, sulphate-reducing bacterial groups, etc., but also different fermentative FTHFS-harbouring bacteria. These fermentative bacteria play important roles in various anaerobic environments and are also of great physiological importance for humans and animals, including *Bacteroidetes, Bifidobacterium, Clostridium, Eubacterium, Faecalibacterium and Lactobacillus* (44–47). AcetoBase thus has great potential to be a representative database for FTHFS-based microbial ecological analyses in varied environments. For instance, the human gut harbours large numbers of acetogenic bacterial communities that are significantly involved in gut physiology, even in the presence of methanogens and sulphate-reducing bacteria (39, 48). Furthermore, members of acetogenic or FTHFS-possessing communities, namely *Clostridium*, *Blautia*, *Eubacterium*, *Eggerthella*, *Prevotella*, *Ruminococcus* and family Lachnospiraceae, are reported to be substantially involved in e.g. dysbiosis of autoimmune disorders, multiple sclerosis, rheumatoid arthritis, systemic lupus erythematosus, cirrhosis, gastrointestinal disorders and Parkinson's disease (49–53). Despite this, only a few studies have examined the FTHFS-harbouring communities in the human gut (54, 55), while any large-scale longitudinal study on this subject is completely lacking.

Conclusions

This paper describes recent updates to AcetoBase in terms of database content, user interface and functionality and presents results from re-analysis, using the updated Acetobase (Version 2), of FTHFS sequence data published in our earlier studies. The updated database content includes additional sequences from new bacterial species and MAGs. This increase in taxonomic diversity is intended to enable FTHFS community profiling and taxonomic annotation with higher accuracy. The updates to the web interface, especially for acetogens and syntrophic acid-oxidizing and sulphate-reducing bacteria, may allow it to serve as a knowledge bank of acetogenic and syntrophic organisms that possess the FTHFS gene. The improved functionality can facilitate searches of the database and retrieval of information on the taxonomic lineage of different bacterial species. The putative taxonomy provided for clone sequences can be of significant help in determining the correct taxonomy, which can differ from the taxonomic associations of published clone sequences, or in future FTHFS clone library-based studies.

The re-analysis of FTHFS sequence data from our previous studies demonstrated and supported the usefulness of the database update in improving the taxonomic annotations resulting from AcetoScan analysis. The variations in taxonomic annotations obtained using the updated AcetoBase Version 2, compared with Version 1, did not change the overall dynamics or interpretations of community profiles. Most of the changes in taxonomic annotations were observed at lower taxonomic levels and among members of the same order or family. The addition of new sequences from MAGs helped improve the identification of FTHFS sequences (higher percentage similarity and lower *e*-value). As amplicon sequence analysis is dependent on reference databases, continuous updates to these databases are needed and the future addition of new sequences may further improve taxonomic annotation. AcetoBase Version 2 contains sequences from acetogens and syntrophic organisms. Detection of this acetogenic and syntrophic composite employing FTHFS amplicon sequencing could open up new avenues in ecology and enable functional studies of microbial interactions in different environments, especially anaerobic bioprocesses and the gut of animals and humans. Ecological analyses and understanding of syntrophic microbial interactions are currently scarce, so longitudinal FTHFS profiling and faster analyses of community dynamics could be an outstanding tool in gaining important insights into the unknown acetogenic and syntrophic microcosm.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

We would like to thank Hans-Henrik Fuxelius for his support and help in the original code modification and developing Version 2 of AcetoBase.

Funding

Swedish University of Agricultural Sciences.

Conflict of interest

None declared.

Data availability

The FTHFS raw sequence data from reactor GR1 analysed in this study and associated with our previous study (4) have been submitted to NCBI SRA (study: SRP336508) under BioProject accession number PRJNA761914. The OTU sequences generated in the re-analysis of data sets 2 and 3 have been submitted to AcetoBase under accession numbers UN_0000023501-UN_0000023887 and UN_0000023888-UN_0000025796, respectively.

Author contributions

A.Si. is involved in the conceptualisation, data curation, methodology, software, visualization and writing—original draft. A.Sc. is involved in the conceptualisation, funding acquisition, resources and writing—review & editing.

References

1. Lovell,C.R. (1994) Development of DNA probes for the detection and identification of acetogenic bacteria. In: Drake HL (ed). *Acetogenesis*, Springer, Boston, MA, pp. 236–253.
2. Lovell,C.R. and Leaphart,A.B. (2004) Community-level analysis: key genes of CO_2-reductive acetogenesis. *Meth. Enzymol.*, **397**, 454–469.
3. Singh,A., Müller,B., Fuxelius,H.-H. *et al.* (2019) AcetoBase: a functional gene repository and database for formyltetrahydrofolate synthetase sequences. *Database*, **2019**, baz142.
4. Singh,A., Nylander,J.A.A., Schnürer,A. *et al.* (2020) High-throughput sequencing and unsupervised analysis of formyltetrahydrofolate synthetase (FTHFS) gene amplicons to estimate acetogenic community structure. *Front Microbiol.*, **11**, 1–13.
5. Drake,H.L. In: Drake HL (ed). (1994) *Acetogenesis. Chapman & Hall Microbiology Series*. Springer, New York, NY, pp. 1–647.
6. Lovell,C.R. and Hui,Y. (1991) Design and testing of a functional group-specific DNA probe for the study of natural populations of acetogenic bacteria. *Appl. Environ. Microbiol.*, **57**, 2602–2609.
7. Leaphart,A.B. and Lovell,C.R. (2001) Recovery and analysis of formyltetrahydrofolate synthetase gene sequences from natural populations of acetogenic bacteria. *Appl. Environ. Microbiol.*, **67**, 1392–1395.
8. Leaphart,A.B., Friez,M.J. and Lovell,C.R. (2003) Formyltetrahydrofolate synthetase sequences from salt marsh plant roots reveal a diversity of acetogenic bacteria and other bacterial functional groups. *Appl. Environ. Microbiol.*, **69**, 693–696.
9. Drake,H.L., Gößner,A.S. and Daniel,S.L. (2008) Old Acetogens, new light. *Ann. N. Y. Acad. Sci.*, **1125**, 100–128.
10. Drake,H.L., Küsel,K. and Matthies,C. (2013) Acetogenic prokaryotes. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds). *The Prokaryotes*. Springer, Berlin, Heidelberg, pp. 3–60.
11. Singh,A., Müller,B. and Schnürer,A. (2021) Profiling temporal dynamics of acetogenic communities in anaerobic digesters using next-generation sequencing and T-RFLP. *Sci. Rep.*, **11**, 13298.
12. Singh,A. *Microbiological Surveillance of Biogas Plants: Focusing on the Acetogenic Community [Internet]*. 2021:12. Swedish University of Agricultural Sciences. http://urn.kb.se/resolve?urn=urn:nbn:se:slu:epsilon-p-110891.
13. Edgar,R. (2016) SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161.
14. Parameswaran,P., Zhang,H., Torres,C.I. *et al.* (2010) Microbial community structure in a biofilm anode fed with a fermentable substrate: the significance of hydrogen scavengers. *Biotechnol. Bioeng.*, **105**, 69–78.
15. Singh,A. (2020) *DupRemover: A Simple Program to Remove Duplicate Sequences from Multi-Fasta File*. ResearchGate/GitHub. https://github.com/abhijeetsingh1704/Duplicate-remover.
16. Singh,A. (2021) *FilterByLength: Filter Fasta Sequences by Length and Count the Sequences in a Multifasta File*. ResearchGate/GitHub. https://github.com/abhijeetsingh1704/FilterByLength.
17. Singh,A. (2019) *Clean_AMINOACID_fasta: Program to Clean AMINO ACID Fasta Sequences and Removes and Illegal Characters in Sequence or Any Non-Natural Amino Acid Residue*. ResearchGate/GitHub. https://github.com/abhijeetsingh1704/clean_AMINOACID_fasta.
18. Singh,A. (2019) *Clean_DNA_fasta: Program to Clean DNA Fasta Sequences and Removes and Illegal Characters in Sequence or Any Non-Natural Residue*. ResearchGate/GitHub. https://github.com/abhijeetsingh1704/clean_DNA_fasta.
19. Rognes,T., Flouri,T., Nichols,B. *et al.* (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4, e2584.
20. Deorowicz,S., Debudaj-Grabysz,A. and Gudyś,A. (2016) FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.*, **6**, 33964.
21. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
22. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. *et al.* (2020) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, **36**, 1925–1927.
23. McSweeney,C., Kang,S., Gagen,E. *et al.* (2009) Recent developments in nucleic acid based techniques for use in rumen manipulation. *Rev. Bras. Zootec.*, **38**, 341–351.
24. Singh,A., Moestedt,J., Berg,A. *et al.* (2021) Microbiological surveillance of biogas plants: targeting acetogenic community. *Front Microbiol.*, **12**, 700256.
25. Campanaro,S., Treu,L., Rodriguez-R,L.M. *et al.* (2020) New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. *Biotechnol. Biofuels*, **13**, 25.
26. Hernsdorf,A.W., Amano,Y., Miyakawa,K. *et al.* (2017) Potential for microbial H_2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. *ISME J.*, **11**, 1915–1929.
27. Coolen,M.J.L. and Orsi,W.D. (2015) The transcriptional response of microbial communities in thawing Alaskan permafrost soils. *Front Microbiol.*, **6**, 197.
28. Saheb-Alam,S., Singh,A., Hermansson,M. *et al.* (2018) Effect of start-up strategies and electrode materials on carbon dioxide reduction on biocathodes. *Appl. Environ. Microbiol.*, **84**, e02242–17.
29. Müller,B., Sun,L., Westerholm,M. *et al.* (2016) Bacterial community composition and *fhs* profiles of low- and high-ammonia biogas digesters reveal novel syntrophic acetate-oxidising bacteria. *Biotechnol. Biofuels*, **9**, 48.
30. Sagheddu,V., Patrone,V., Miragoli,F. *et al.* (2017) Abundance and diversity of hydrogenotrophic microorganisms in the infant gut before the weaning period assessed by denaturing gradient gel electrophoresis and quantitative PCR. *Front. Nutr.*, **4**, 29.
31. Breznak,J. (1990) Microbial H_2/CO_2 acetogenesis in animal guts: nature and nutritional significance. *FEMS Microbiol. Lett.*, **87**, 309–313.
32. Küsel,K. and Drake,H.L. (1994) Acetate synthesis in soil from a Bavarian beech forest. *Appl. Environ. Microbiol.*, **60**, 1370–1373.
33. Zhang,H., Parameswaran,P., Badalamenti,J. *et al.* (2011) Integrating high-throughput pyrosequencing and quantitative real-time PCR to analyze complex microbial communities. *Methods Mol. Biol.*, **733**, 107–128.
34. Martinez,M.A., Woodcroft,B.J., Ignacio Espinoza,J.C. *et al.* (2019) Discovery and ecogenomic context of a global Caldiserica-related phylum active in thawing permafrost, Candidatus Cryosericota phylum nov., Ca. Cryosericia class nov., Ca. Cryosericales ord. nov., Ca. Cryosericaceae fam. nov., comprising the four species Cryosericum septentrionale gen. nov. sp. nov., Ca. C. hinesii sp. nov., Ca. C. odellii sp. nov., Ca. C. terrychapinii sp. nov. *Syst. Appl. Microbiol.*, **42**, 54–66.
35. Wolin,M.J. and Miller,T.L. (1994) Acetogenesis from CO_2 in the human colonic ecosystem. In: Drake HL (ed). *Acetogenesis*, Chapman & Hall Microbiology Series, Springer, Boston, MA, pp. 365–385.
36. Das,A. and Ljungdahl,L.G. (2003) Electron-transport system in acetogens. In: Ljungdahl LG, Adams MW, Barton LL, Ferry JG,

Johnson MK, (eds). *Biochemistry and physiology of anaerobic bacteria*, pp. 191–204.
37. Hügler,M. and Sievert,S.M. (2011) Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Ann. Rev. Mar. Sci.*, **3**, 261–289.
38. Yang,C. (2018) Acetogen communities in the gut of herbivores and their potential role in syngas fermentation. *Fermentation*, **4**, 40.
39. Rey,F.E., Faith,J.J., Bain,J. *et al.* (2010) Dissecting the in vivo metabolic potential of two human gut acetogens. *J. Biol. Chem.*, **285**, 22082–22090.
40. Lechtenfeld,M., Heine,J., Sameith,J. *et al.* (2018) Glycine betaine metabolism in the acetogenic bacterium *Acetobacterium woodii*. *Environ. Microbiol.*, **20**, 4512–4525.
41. Kountz,D.J., Behrman,E.J., Zhang,L. *et al.* (2020) MtcB, a member of the MttB superfamily from the human gut acetogen *Eubacterium limosum*, is a cobalamin-dependent carnitine demethylase. *J. Biol. Chem.*, **295**, 11971–11981.
42. Picking,J.W., Behrman,E.J., Zhang,L. *et al.* (2019) MtpB, a member of the MttB superfamily from the human intestinal acetogen *Eubacterium limosum*, catalyzes proline betaine demethylation. *J. Biol. Chem.*, **294**, 13697–13707.
43. Wen,L.-L., Zhang,Y., Pan,Y.-W. *et al.* (2015) The roles of methanogens and acetogens in dechlorination of trichloroethene using different electron donors. *Environ. Sci. Pollut. Res.*, **22**, 19039–19047.
44. Collado,M.C., D'Auria,G., Mira,A. *et al.* (2013) Human microbiome and diseases: a metagenomic approach. In: Watson RR, Preedy VRBT-BF as DI for L and GD, (eds). *Bioactive Food as Dietary Interventions for Liver and Gastrointestinal Disease*, pp. 235–249.
45. Panasevich,M.R., Kerr,K.R., Dilger,R.N. *et al.* (2015) Modulation of the faecal microbiome of healthy adult dogs by inclusion of potato fibre in the diet. *Br. J. Nutr.*, **113**, 125–133.
46. Induri,S.N.R., Kansara,P., Thomas,S.C. *et al.* (2022) The gut microbiome, metformin, and aging. *Annu. Rev. Pharmacol. Toxicol.*, **62**, 85–108.
47. Kanauchi,O., Fujiyama,Y., Mitsuyama,K. *et al.* (1999) Increased growth of *Bifidobacterium* and *Eubacterium* by germinated barley foodstuff, accompanied by enhanced butyrate production in healthy volunteers. *Int. J. Mol. Med.*, **3**, 175–179.
48. Dore,J. (1995) Enumeration of H_2-utilizing methanogenic archaea, acetogenic and sulfate-reducing bacteria from human feces. *FEMS Microbiol. Ecol.*, **17**, 279–284.
49. de Oliveira,G.L.V. (2019) The gut microbiome in autoimmune diseases. In: Faintuch J, Faintuch S, (eds). *Microbiome and Metabolome in Diagnosis, Therapy, and Other Strategic Applications*, pp. 325–332.
50. Eslamparast,T., Eghtesad,S., Hekmatdoost,A. *et al.* (2013) Probiotics and nonalcoholic fatty liver disease. *Middle East J. Dig. Dis.*, **5**, 129–136.
51. Keshavarzian,A., Engen,P., Bonvegna,S. *et al.* (2020) The gut microbiome in Parkinson's disease: a culprit or a bystander? *Prog. Brain Res.*, **252**, 357–450.
52. Raghavendra,P. and Pullaiah,T. (2018) Pathogen identification using novel sequencing methods. *Adv. Cell Mol. Diagn.*, 161–202.
53. Floch,N. (2017) The influence of microbiota on mechanisms of bariatric surgery. In: Floch MH, Ringel Y, Allan Walker WBT-TM in GP, (eds). *The Microbiota in Gastrointestinal Pathophysiology*, pp. 267–281.
54. Ohashi,Y., Igarashi,T., Kumazawa,F. *et al.* (2007) Analysis of Acetogenic Bacteria in human feces with formyltetrahydrofolate synthetase sequences. *Biosci. Microflora*, **26**, 37–40.
55. Ohashi,Y., Andou,A., Kanaya,M. *et al.* (2009) Acetogenic Bacteria mainly contribute to the disposal of hydrogen in the colon of healthy Japanese. *Biosci. Microflora*, **28**, 17–19.