MICROBIAL GENOMICS

RESEARCH ARTICLE
Pidcock *et al.*, *Microbial Genomics* 2021;7:000638
DOI 10.1099/mgen.0.000638

Phylogenetic systematics of *Butyrivibrio* and *Pseudobutyrivibrio* genomes illustrate vast taxonomic diversity, open genomes and an abundance of carbohydrate-active enzyme family isoforms

Sara E. Pidcock[1], Timofey Skvortsov[2], Fernanda G. Santos[1], Stephen J. Courtney[1], Karen Sui-Ting[1], Christopher J. Creevey[1] and Sharon A. Huws[1,*]

Abstract

Butyrivibrio and *Pseudobutyrivibrio* dominate in anaerobic gastrointestinal microbiomes, particularly the rumen, where they play a key role in harvesting dietary energy. Within these genera, five rumen species have been classified (*Butyrivibrio fibrisolvens*, *Butyrivibrio hungatei*, *Butyrivibrio proteoclasticus*, *Pseudobutyrivibrio ruminis* and *Pseudobutyrivibrio xylanivorans*) and more recently an additional *Butyrivibrio* sp. group was added. Given the recent increase in available genomes, we re-investigated the phylogenetic systematics and evolution of *Butyrivibrio* and *Pseudobutyrivibrio*. Across 71 genomes, we show using 16S rDNA and 40 gene marker phylogenetic trees that the current six species designations (*P. ruminis*, *P. xylanivorans*, *B. fibrisolvens*, *Butyrivibrio* sp., *B. hungatei* and *B. proteclasticus*) are found. However, pangenome analysis showed vast genomic variation and a high abundance of accessory genes (91.50–99.34%), compared with core genes (0.66–8.50%), within these six taxonomic groups, suggesting incorrectly assigned taxonomy. Subsequent pangenome accessory genomes under varying core gene cut-offs (%) and average nucleotide identity (ANI) analysis suggest the existence of 42 species within 32 genera. Pangenome analysis of those that still group within *B. fibrisolvens*, *B. hungatei* and *P. ruminis*, based on revised ANI phylogeny, also showed possession of very open genomes, illustrating the diversity that exists even within these groups. All strains of both *Butyrivibrio* and *Pseudobutyrivibrio* also shared a broad range of clusters of orthologous genes (COGs) (870), indicating recent evolution from a common ancestor. We also demonstrate that the carbohydrate-active enzymes (CAZymes) predominantly belong to glycosyl hydrolase (GH)2, 3, 5, 13 and 43, with numerous within family isoforms apparent, likely facilitating metabolic plasticity and resilience under dietary perturbations. This study provides a major advancement in our functional and evolutionary understanding of these important anaerobic bacteria.

DATA SUMMARY

Pseudobutyrivibrio xylanivorans MZ8 was genome sequenced in this study and the sequence submitted to GenBank (BioProject number PRJNA563299) (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA563299). Supplementary Material can be found with the online version of this article and on FigShare (https://doi.org/10.6084/m9.figshare.14785131.v1).

INTRODUCTION

The definition of 'species' in bacteria or archaea is contentious, with some believing that the search for a single, natural way to divide bacteria into species is futile [1, 2]. Historically, the most important characteristics in terms of taxonomic markers were morphology, growth requirements and pathogenic potential. At the beginning of the 20th century, more biochemical and physiological markers were added to this list, followed by chemotaxonomy, numerical taxonomy and

Received 02 January 2021; Accepted 22 June 2021; Published 04 October 2021
Author affiliations: [1]School of Biological Sciences and Institute for Global Food Security, 19 Chlorine Gardens, Queen's University Belfast, Belfast BT9 5DL, UK; [2]School of Pharmacy, Medical Biology Centre, 97 Lisburn Road, Queen's University Belfast, Belfast BT9 7BL, UK.
***Correspondence:** Sharon A. Huws, s.huws@qub.ac.uk
Keywords: *Butyrivibrio*; evolution; pangenome; *Pseudobutyrivibrio*; rumen; taxonomy.
Abbreviations: AGE, accessory genomic element; ANI, average nucleotide identity; CAZyme, carbohydrate-active enzyme; COG, cluster of orthologous genes; GH, glycosyl hydrolase; iTOL, Interactive Tree of Life; RPKM, reads per kilobase of transcript per million mapped reads.
The GenBank BioProject accession number for the genome sequence of *Pseudobutyrivibrio xylanivorans* MZ8 is PRJNA563299.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Two supplementary tables, fourteen supplementary figures and two supplementary Excel datasets are available with the online version of this article and on FigShare (https://doi.org/10.6084/m9.figshare.14785131.v1).

DNA–DNA hybridization in the mid-late 20th century. More recently, we have also used genotypic analyses, multilocus sequence analyses, average nucleotide identity (ANI), whole-genome analyses, etc. [3, 4]. 16S rDNA became a popular metric in the 1980s, with organisms sharing greater than 97% 16S rDNA being classified as a single species [5]. This was further developed to whole-genome alignments [6] and phylogenetic clustering [7], both facing scrutiny for their seemingly arbitrary cut-off values [8]. 16S rDNA was also criticised on the basis that only a single gene is used as a point of comparison [9], prompting the development of sets of universal marker genes, which were proposed to form a more resolved phylogeny, but these constructions were far more computationally demanding [10]. It is inevitable that some degree of subjectivity is seen with respect to taxonomy, and consequently the same group of organisms can be sorted and arranged in many different ways [4, 11]. More recently, pangenomic analyses (those which look at shared core genes, accessory genes that confer variability, and the combination of these as the pangenome) have also been suggested as potential methods for defining bacterial species [12–14].

The rumen microbiome is taxonomically ambiguous, with horizontal gene transfer being rife due to the intense proximity that it provides [15]. Consequently, our understanding of both the taxonomy and function of the constituent microbes remains vague as it is in constant flux. Recently, our understanding has been enhanced through the Hungate collection [16], which comprise 501 rumen microbial genomes. Recent studies show that a core microbiome including *Prevotella*, *Butyrivibrio* and *Ruminococcus* can be found in ruminants globally [17]. Nonetheless, the taxonomy of *Butyrivibrio* and *Pseudobutyrivibrio* remains a topic of debate.

Butyrivibrio were first described in 1956 using classical morphological and biochemical taxonomy; they are motile, Gram-positive, slightly curved rods that produce large amounts of butyric acid via glucose fermentation [18–20]. It was noted upon discovery that extensive variation within the genus may lead to difficulties in defining species–specific patterns [18, 19, 21]. They were called *Butyrivibrio fibrisolvens*, after their importance in the digestion of fibre in ruminant feed [18, 19] via carbohydrate-active enzymes (CAZymes) [22]. New strains were routinely classified as *B. fibrisolvens* despite morphological and genetic diversity [23] until 1976, when *Butyrivibrio crossotus*, a predominantly human isolate, was first described [24]. Thereafter, four *Butyrivibrio* species were defined: *B. fibrisolvens*, *B. crossotus*, *Butyrivibrio hungatei* and *Butyrivibrio proteoclasticus* (Fig. 1) [19, 23–25]. In 2008, *B. proteoclasticus* was reclassified, originally being *Clostridium proteoclasticum*, based on phylogenetic placement, DNA G+C content and physiological traits [25]. In 1996, a bacterium was isolated that resembled *B. fibrisolvens*, but varied sufficiently based on 16S rDNA, G+C content and cellular fatty acid content, and was named *Pseudobutyrivibrio ruminis* [26]. Later in 2003, *Pseudobutyrivibrio xylanivorans* was classified based upon fermentation characteristics, DNA G+C content and 16S rDNA dissimilarity to *Butyrivibrio* spp. [23]. Despite these additional species being named, *Butyrivibrio* and *Pseudobutyrivibrio* still possess untapped phylogenetic and genetic diversity with a sixth undefined *Butyrivibrio* sp. grouping identified recently [27].

Impact Statement

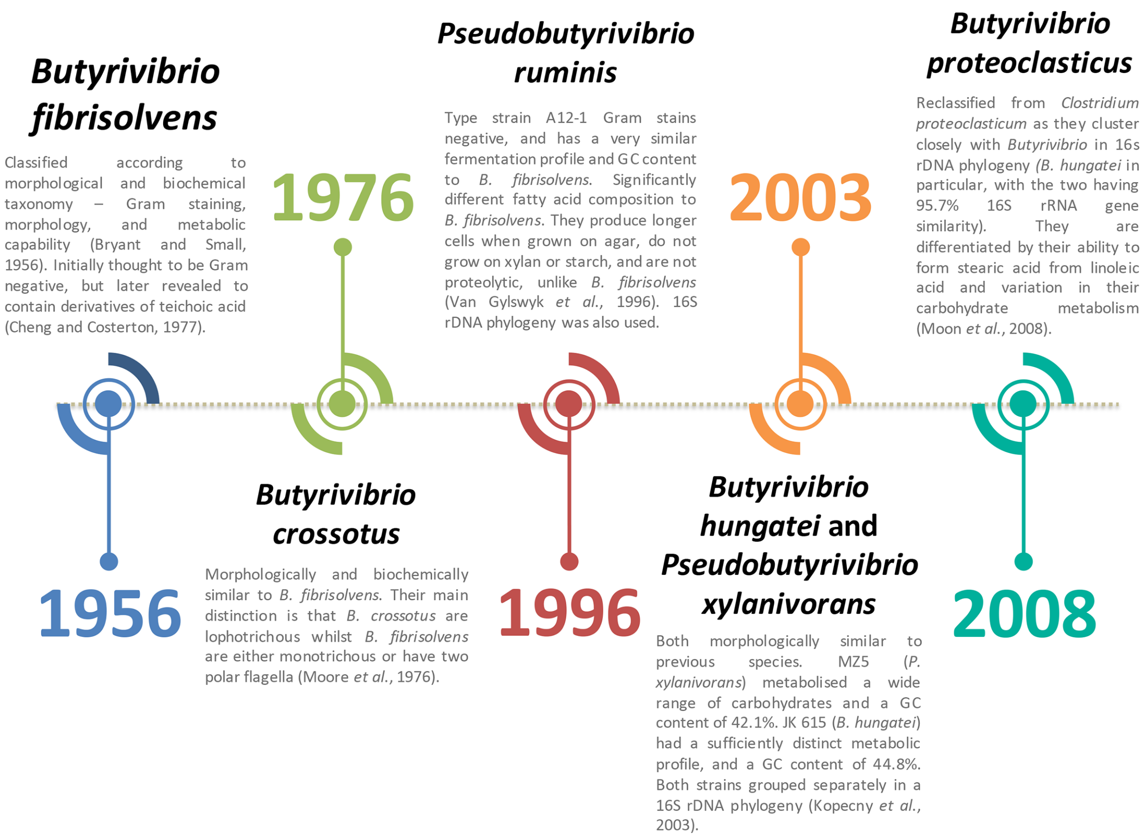
Previous studies have suggested that immense intra- and inter-genetic variation lies within the genera *Butyrivibrio* and *Pseudobutyrivibrio*. Whether this is an artefact of inconsistent taxonomic approaches or a legitimate occurrence within the genera is not known. As such, their current taxonomic designations should be reviewed using current methodologies. This is also particularly timely as recently there has been a major increase in deposited available genomes for these genera. Consequently, this study reviews the taxonomy of *Butyrivibrio* and *Pseudobutyrivibrio* using computational approaches [including pangenomics, average nucleotide identity (ANI) and gene orthology] to determine the validity of their current taxonomy. We found that the current six species of *Butyrivibrio* and *Pseudobutyrivibrio* underrepresent the true taxonomic diversity and suggest based on ANI and coverage that the 71 genomes used in this study constitute 42 species within 32 genera, with those that still group within *Butyrivibrio fibrisolvens*, *Butyrivibrio hungatei* and *Pseudobutyrivibrio ruminis* based on revised ANI phylogeny showing very open genomes. As such, we suggest re-evaluating the species and genus designations of the strains included in this study. Despite genetic dissimilarity, all strains appear to maintain a similar broad functional profile within the rumen and share a broad range of clusters of orthologous genes (COGs), indicating fairly recent evolution from a common ancestor. Strains also possess an abundance of glycosyl hydrolase isoforms, which may afford them greater metabolic plasticity *in vivo*.

The aims of this study were to re-investigate phylogeny, gene-level functional divergence and evolution in predominantly ruminal *Butyrivibrio* and *Pseudobutyrivibrio* using all publicly available genomes from pure cultures. This study also investigated gene-centric evolution in the ruminal *Butyrivibrio* and *Pseudobutyrivibrio* in relation to their pangenomes. In parallel, we also classified their carbohydrate degrading capacity. Many of the bacterial genomes in the recently expanded Hungate collection are from the genera *Butyrivibrio* and *Pseudobutyrivibrio*, which makes this study timely, enabling a paradigm shift in our fundamental understanding of these genera.

METHODS

Genomes used in this study

Seventy-one genomes of *Butyrivibrio*/*Pseudobutyrivibrio* isolates were obtained from the Hungate collection (Joint



1956

Butyrivibrio fibrisolvens

Classified according to morphological and biochemical taxonomy – Gram staining, morphology, and metabolic capability (Bryant and Small, 1956). Initially thought to be Gram negative, but later revealed to contain derivatives of teichoic acid (Cheng and Costerton, 1977).

1976

Butyrivibrio crossotus

Morphologically and biochemically similar to *B. fibrisolvens*. Their main distinction is that *B. crossotus* are lophotrichous whilst *B. fibrisolvens* are either monotrichous or have two polar flagella (Moore *et al.*, 1976).

1996

Pseudobutyrivibrio ruminis

Type strain A12-1 Gram stains negative, and has a very similar fermentation profile and GC content to *B. fibrisolvens*. Significantly different fatty acid composition to *B. fibrisolvens*. They produce longer cells when grown on agar, do not grow on xylan or starch, and are not proteolytic, unlike *B. fibrisolvens* (Van Gylswyk *et al.*, 1996). 16S rDNA phylogeny was also used.

2003

Butyrivibrio hungatei* and *Pseudobutyrivibrio xylanivorans

Both morphologically similar to previous species. MZ5 (*P. xylanivorans*) metabolised a wide range of carbohydrates and a GC content of 42.1%. JK 615 (*B. hungatei*) had a sufficiently distinct metabolic profile, and a GC content of 44.8%. Both strains grouped separately in a 16S rDNA phylogeny (Kopecny *et al.*, 2003).

2008

Butyrivibrio proteoclasticus

Reclassified from *Clostridium proteoclasticum* as they cluster closely with *Butyrivibrio* in 16s rDNA phylogeny (*B. hungatei* in particular, with the two having 95.7% 16S rRNA gene similarity). They are differentiated by their ability to form stearic acid from linoleic acid and variation in their carbohydrate metabolism (Moon *et al.*, 2008).

Fig. 1. Chronological identification and classification of *Butyrivibrio* and *Pseudobutyrivibrio*.

Genome Institute) [16] and including one additional strain, *P. xylanivorans* MZ8 (obtained from the Rowett Research Institute, University of Aberdeen, Aberdeen, UK), genome sequenced (Table S1, available with the online version of this article) by ourselves using MicrobesNG (https://microbesng.uk/) (sequenced on the Illumina HiSeq 2500 platform, using 2×250 bp paired-end reads and with 30× coverage) and submitted to GenBank (BioProject number PRJNA563299). The data were put through MicrobesNG's standard analysis pipeline, which included strain identification by Kraken [28], *de novo* assembly of the reads by SPAdes [29] and Prokka V1.12 annotation [30]. However, we re-annotated all 71 genomes using Prokka v1.12 via the Galaxy platform [31] with a similarity E value cut-off of 1×10^{-6} to ensure all were annotated consistently.

Phylogeny

16S rDNA sequences, obtained via Prokka annotations of the genomes, were aligned using the Aligner pipeline of the Ribosomal Database Project (RDP), release 11, update 5, September 30 2016 [32], and an approximately-maximum-likelihood phylogenetic tree with 1000 bootstrap repetitions was reconstructed using FastTree v2.1.10 [33]. An additional tree was reconstructed using 40 gene markers as per Wu and Eisen [34], and Creevey *et al.* [35]. Both trees were visualized by the Interactive Tree of Life (iTOL), changelog version 3.5.2 [36]. *Clostridium beijerinckii* NCIMB 8052 and *Lactobacillus acidophilus* NCFM were used as outgroups.

ANI was calculated using the PyANI script (available at https://github.com/widdowquinn/pyani/tree/version_0_2) [37]. Input sequences were in FASTA format and were aligned using MUMmer (NUCmer). The comparisons were visualized by selecting for heatmap and dendrogram output.

Pangenomics

Pangenomic analysis was carried out according to the classical six species taxonomy (*B. fibrisolvens*, *B. hungatei*, *B. proteoclasticus*, *Butyrivibrio* sp., *P. ruminis* and *P. xylanivorans*) to further define their phylogeny. Pangenomics was also used to define core and accessory genes within ANI-defined groupings containing at least four genome representatives. Core and accessory genomic fragments were identified from the Prokka annotated genomic sequences (.ffn files) using Spine v0.3.1 (http://vfsmspineagent.fsm.northwestern.edu/index_age.html) [38]. A range of other defined parameters (70–100% similarity and present in 50–100% of genomes) were evaluated, with 90% [38–40], and the default value of 100% core [41–43] definitions, using the default value of 85% identity. Core values of 100 and 85% identity were subsequently used

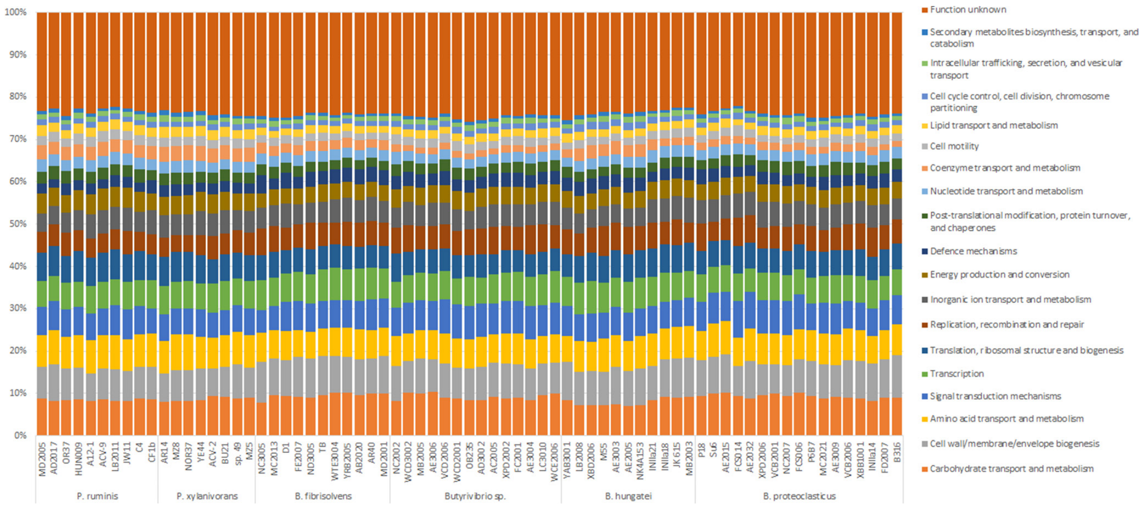


Fig. 2. Functional annotation of the 71 *Butyrivibrio* and *Pseudobutyrivibrio* genomes used in this study. Gene functionality is sorted by colour, as indicated in the key. Annotation was performed using EggNOG [44] (http://eggnogdb.embl.de/#/app/emapper).

as the pangenomics parameters. Accessory elements were visualized using ClustAGE v0.8 and ClustAGE plot [38] for each group. The minimum accessory genomic element (AGE) size to represent in the ClustAGE plot was set to 1500 bp.

Core and accessory FASTA files for individual strains were combined into the core genome and the accessory genome for their respective taxa based upon classical taxonomy and the ANI-refined taxonomy for the three groups with >4 genome representatives. We then prospected the *B. fibrisolvens* classical and ANI-defined core and accessory gene data for presence of the core orthologous genes used to conduct the 40 marker tree [11], using BLASTX and cut-off identity of 80%. This was completed as these genes should be classified as core and as such were used as an estimate of the precision of core and accessory gene definitions for classical and ANI-refined taxonomy. Divergent bacteria will result in some core genes being incorrectly assigned as accessory. Core and accessory genes for the three ANI-refined groupings were uploaded to EggNOG [44] to determine genomic subsystem annotation. A stacked histogram was then made using these data to compare core and accessory functionality on a taxon level. Core and accessory gene G+C mol% content were also calculated from the Spine statistics files.

Gene evolution

Putative gene orthology was determined using OrthAgogue [45] using the amino acid sequences of Prokka-annotated genes (.faa) from all 71 genomes. OrthAgogue was run with the parameters '-b -e 6', which set the E value cut-off to 10^{-6} and forced OrthoMCL [46] emulation. We also ran OrthAgogue with an E value cut-off of 10^{-5} but the results were an E value cut-off to 10^{-6}, which we decided to use for downstream analysis. All other OrthAgogue parameters were default. The clusters of orthologous genes (COGs) identified by OrthAgogue were turned into binary data, after which the lists of COGs were uploaded to UpSet [47] to visualize the intersections. It should be noted that output from OrthAgogue does not represent 1:1 gene orthology, rather it lists of genes that are part of clusters of orthologous genes.

CAZymes

CAZymes were identified using the dbCAN metaserver [48] and annotated using the 'mRNAs/CDSs/Metagenomes or short DNA seqs' option, running HMMER with default setting of E value $<1\times10^{-15}$, coverage >0.35. Sequences were extracted using SAMtools v1.9 and each glycosyl hydrolase (GH) family was aligned using the ClustalOmega online server [49]. Trees were inferred by maximum likelihood using IQ-TREE v1.6.10, and visualized using iTOL. Homologue-based annotations were derived from dbCAN. Stacked histograms were produced for the most abundant GH families; in the case of the histogram with all 71 genomes displayed, and the histogram showing all six taxa, only GH families with over 100 total instances across all genomes were displayed. For the taxon-specific histograms, this number was 10. Pairwise per cent identity between GH sequences was determined by uploading amino acid sequences to the ClustalOmega online server, which produced a pairwise identity matrix.

Metatranscriptome analysis

To check expression of the identified CAZyme isoforms in the rumen and *in vivo*, 20 publicly available metatranscriptomic datasets were taken from the National Center for Biotechnology Information Sequence Read Archive, under the accession number SRA075938. Datasets were composed of 150 bp paired-end reads from the Illumina HiSeq 2000 sequencer. FASTQ files were processed with

Table 1. Phylogeny of the 71 *Butyrivibrio* and *Pseudobutyrivibrio* strains used in this study, according to 16S rDNA, 40 marker tree and ANI analysis

Group	16S rDNA tree	40 marker tree	ANI
B. fibrisolvens	AB2020, AR40, D1, FE2007, MC2013, MC2021, MD2001, NC3005, ND3005, TB, WTE3004, YRB2005	AB2020, AR40, D1, FE2007, MC2013, MD2001, NC3005, ND3005, TB, WTE3004, YRB2005	AB2020, AR40, FE2007, MD2001, ND3005, TB, WTE3004, YRB2005
B. hungatei	AE2005, AE3003, INlla18, JK615, LB2008, M55, MB2003, NK4A153, VCB2006, XBD2006	AE2005, AE3003, INlla18, INlla21, JK615, LB2008, M55, MB2003, NK4A153, XBD2006, YAB3001	AE2005, AE3003, M55, NK4A153
B. proteoclasticus	AE2015, AE2032, AE3009, B316, FCS006, FCS014, FD2007, INlla14, INlla21, NC2007, P18, P6B7, Su6, VCB2001, XBB1001, XPD2006, YAB3001	AE2015, AE2032, AE3009, B316, FCS006, FCS014, FD2007, INlla14, MC2021, NC2007, P18, P6B7, Su6, VCB2001, VCB2006, XBB1001, XPD2006	
Butyrivibrio sp.	AC2005, AD3002, AE3004, AE3006, FC2001, LC3010, MB2005, OB235, VCD2006, WCD2001, WCD3002, WCE2006, XPD2002	AC2005, AD3002, AE3004, AE3006, FC2001, LC3010, MB2005, NC2002, OB235, VCD2006, WCD2001, WCD3002, WCE2006, XPD2002	
P. ruminis	A12-1, ACV-9, AD2017, C4, CF1b, HUN009, JW11, LB2011, NC2002, YE44	A12-1, ACV-9, AD2017, C4, CF1b, HUN009, JW11, LB2011, MD2005, OR37	A12-1, ACV-9, C4, HUN009, JW11, LB2011
P. xylanivorans	49, ACV-2, AR14, Bu21, MD2005, Mz5, Mz8, NOR37, OR37	49, ACV-2, AR14, Bu21, Mz5, Mz8, NOR37, YE44	

multiqc [50] and reads were trimmed from 150 to 110 bp using trimmomatic software v0.36 [51]. Reads were aligned to the Hungate rumen genome dataset using bowtie2 v2.3.0 [52] using the settings '--very-sensitive-local', allowing soft trimming, and a relaxed alignment; and '-k 497'. This produced SAM files, which were converted to BAM files using SAMtools v1.9 [53]. SAMtools v1.9 was used to filter all and the best alignment position for each read using the flag option '-F 260'. For each BAM file, FeatureCounts (from the subread package v2.0.0) [54] was used to calculate the number of reads that align within the boundaries of every predicted gene in the Hungate genomes. Read counts were converted into RPKM (reads per kilobase of transcript per million mapped reads) values. RPKM values of the CAZyme gene haplotypes were extracted from the entire expression count table. If a gene was found in a

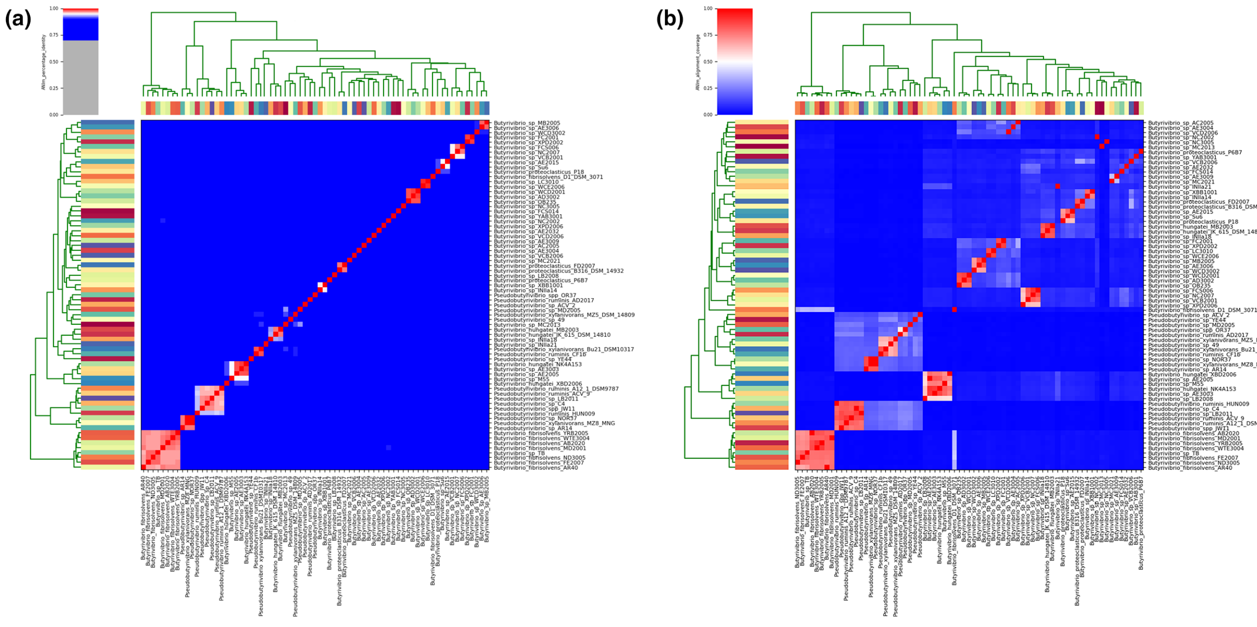


Fig. 3. ANI comparison of 71 strains of *Butyrivibrio* and *Pseudobutyrivibrio* using pyani.py (https://github.com/widdowquinn/pyani/tree/version_0_2) with the MUMmer alignment option (a) (cells in the heat map that are coloured red have >95% sequence similarity, whilst blue cells have <95% similarity, and as nucleotide identity reaches 95% the cells are coloured white). Alignment coverage of all strains using pyani.py with the MUMmer alignment option (b) (cells in the heat map that are coloured red have >50% coverage, whilst blue cells have <50% similarity, and as nucleotide identity reaches 95% the cells are coloured white).

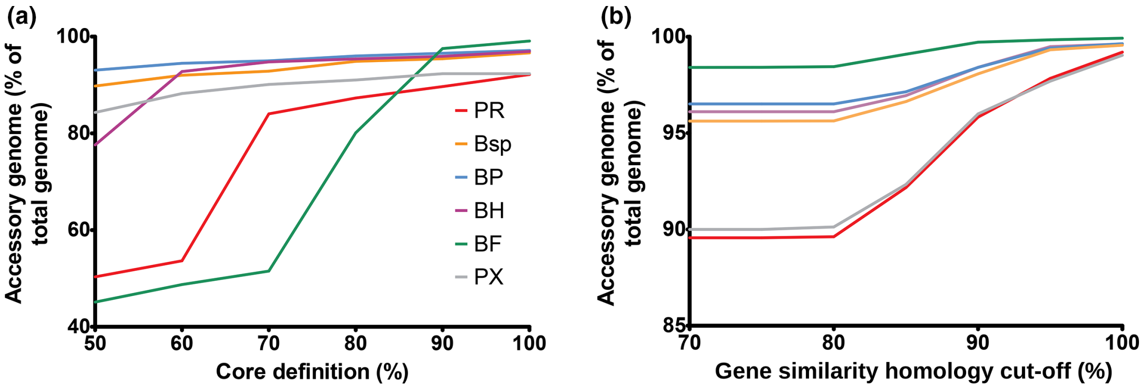


Fig. 4. Effect of core definition percentage (a) and gene similarity homology (b) cut-off percentage on the pangenome composition of current *Butyrivibrio* and *Pseudobutyrivibrio* taxa as denoted by coloured lines: green being *B. fibrisolvens* (BF), purple being *B. hungatei* (BH), blue being *B. proteoclasticus* (BP), orange being *Butyrivibrio* sp. (Bsp), red being *P. ruminis* (PR) and grey being *P. xylanivorans* (PX).

metatranscriptome (expressed to any degree) then it was visualized on the iTOL GH trees.

RESULTS

Bacterial function

Functional analysis was completed for all 71 ruminal *Butyrivibrio* and *Pseudobutyrivibrio* genomes used in this study using EggNOG-mapper (Fig. 2, Table S1). Overall, strains have predominant functions relating to carbohydrate transport and metabolism (mean 8.91%), and cell wall, membrane and envelope genesis (mean 8.05%). These stay constant across all strains (0.80 and 0.78% SD). This general consistency can also be seen when the functional categories are compared at a genus level (Fig. S1).

Phylogenetic analysis using 16S rDNA, 40 marker trees and ANI

Phylogenetic trees based on 16S rDNA and 40 marker genes were reconstructed to examine phylogenetic relatedness (Figs S2 and S3, respectively). Both phylogenies form six groups that correspond to the classical taxonomy, including the additional group comprised entirely of strains allocated to *Butyrivibrio* but with no species designation (Table 1). It should be noted that *B. hungatei* forms a paraphyletic clade. Similar groupings can be seen again in the G+C-content-based scatter plot (Fig. S4), with the classical taxa *P. ruminis* and *P. xylanivorans* grouping very closely, as well as *B. hungatei* and *B. proteoclasticus*. However, ANI data showed that most strains have less than 95% pairwise nucleotide identity (Fig. 3a), with small groups being above this

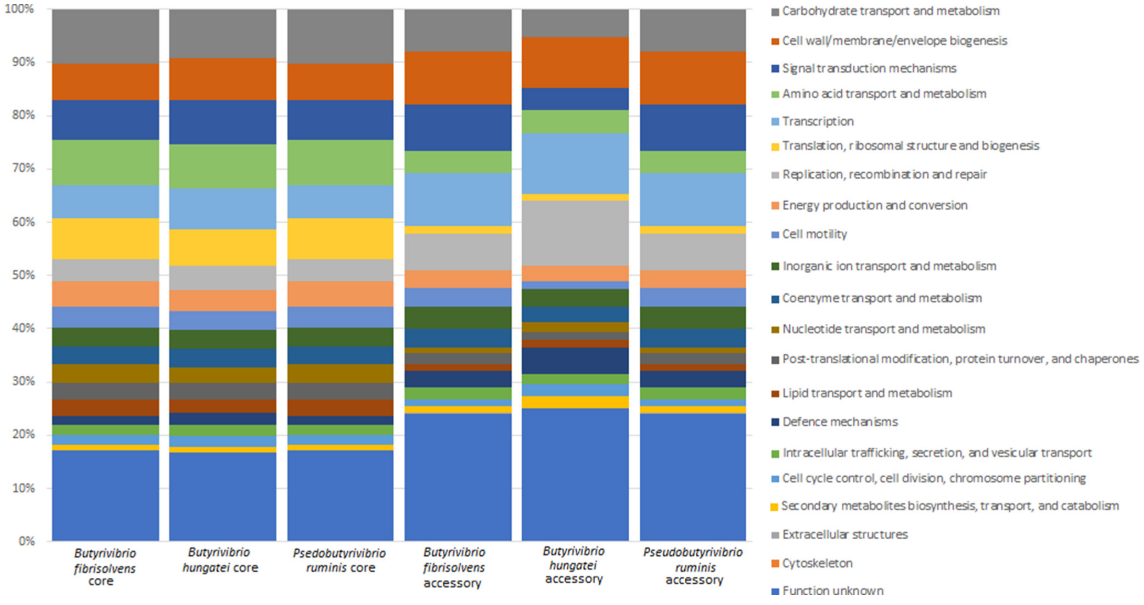


Fig. 5. Functional annotations of the core and accessory genomes of the ANI-defined *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups.

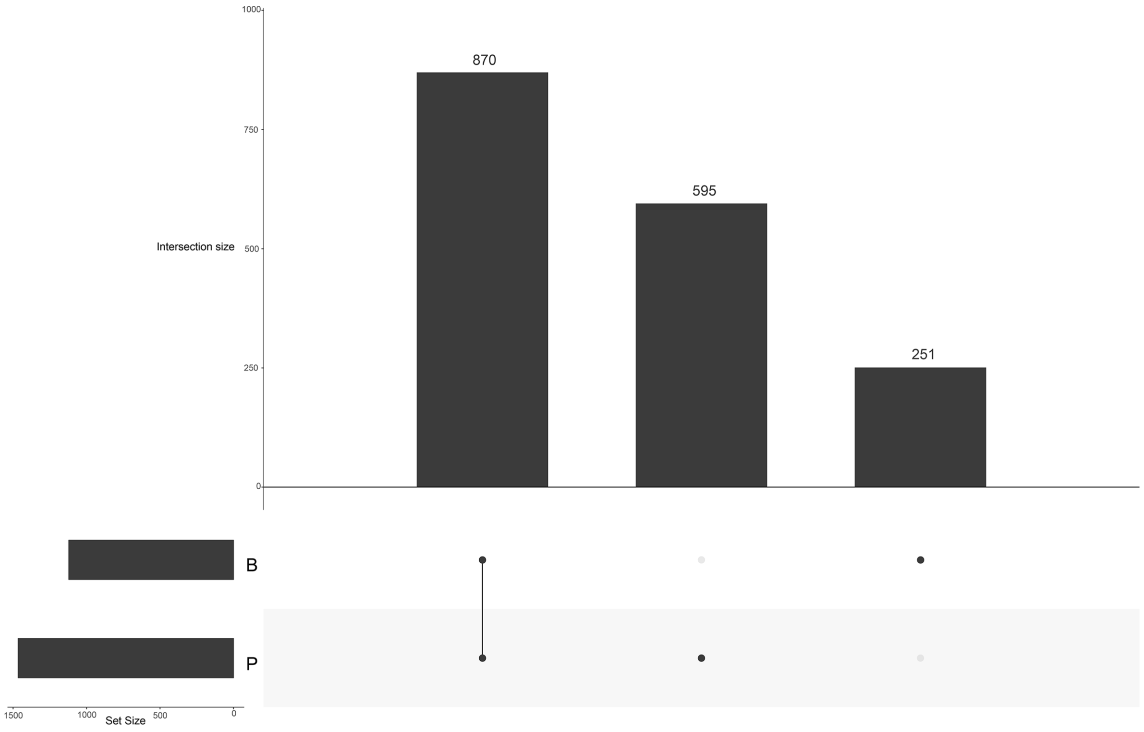


Fig. 6. UpSet plot showing orthologous gene cluster intersections across the two genera, *Butyrivibrio* (B) and *Pseudobutyrivibrio* (P). Intersections are denoted by the corresponding bar.

threshold. Many of these strains have <50% genome coverage, meaning that they align to each other for less than 50% of their genome length (Fig. 3b). Indeed, ANI data suggests the existence of 32 genera and 42 species (Fig. 3a, b).

Pangenomics

Pangenomes were investigated using Spine software and a range of Spine cut-off parameters (core definition per cent and per cent nucleotide identity) were tested (Fig. 4a, b) . These data show that, as expected, the more stringent the core definition is (i.e. the more genomes of the total population that the gene must be in to be considered core), the more genes are considered to be accessory. These data also show the appearance of 'bumps' in the line graphs as parameters change, suggesting the presence of further taxonomical diversity within the classical clades (Fig. 4a, b). For downstream analysis, the default parameter of 100% core definition was used as well as the default 85% gene nucleotide identity cut-off based on choosing a mid-point that was both stringent but allowed some flexibility to account for the highly divergent nature of the bacteria (Table S1).

Using the classical taxa at a 100% core definition, *B. fibrisolvens* has a mean core G+C content of 44.94mol%, and an accessory G+C content of 40.57 mol%. *B. hungatei* has 44.73 and 41.04mol% for mean core and accessory G+C contents, respectively; *B. proteoclasticus* 45.23 and 42.70mol%, respectively; *Butyrivibrio* sp. 45.14 and 41.58mol%, respectively; *P. ruminis* 43.39 and 39.06mol%, respectively; and *P. xylanivorans* 43.69 and 39.12mol%, respectively. *Pseudobutyrivibrio* strains had a greater difference than *Butyrivibrio* strains, with a mean difference of 4.32mol% compared with 3.38mol%. The greatest difference was seen in strains of *P. xylanivorans*, with a mean 4.57mol% difference in G+C content. Core genes across each taxon appear to have a higher G+C mol% than their respective accessory genomes (with a mean of 44.57mol% in the core genome, and 40.95mol% in the accessory at a core definition of 90%). This pattern can also be seen in the three main groups obtained using ANI (existing within *B. fibrisolvens*, *B. hungatei* and *P. ruminis* in the classical taxonomy), with the average G+C mol% for the three accessory pangenomes being 38.0mol%, and the core being 41.55mol% (Table S1). When based on the ANI groups (Table 1), core genomes were larger than that of classical taxa, comprising means of 59.62, 68.50, and 65.23mol% for groups 1, 2 and 3 (*B. fibrisolvens*, *B. hungatei* and *P. ruminis*, respectively), although still illustrating very open genomes.

ClustAGE plots for classical taxa show high levels of genomic dissimilarity, with large numbers of AGEs being distributed throughout the genome (Figs S5–S10). The plot for *B. fibrisolvens* as a classical taxon shows gene fragments being absent in many genomes in places. This is particularly clear on occasion, for example, at the 750 kbp mark, with only 4 genomes out of 11 (AB2020, MC2013, NC3005 and D1) possessing an AGE here (Fig. S5). The ClustAGE plots for the PyANI groupings show sparser AGE fragments (due to a smaller accessory genome) but they appear to be similarly distributed (Figs S11–S13).

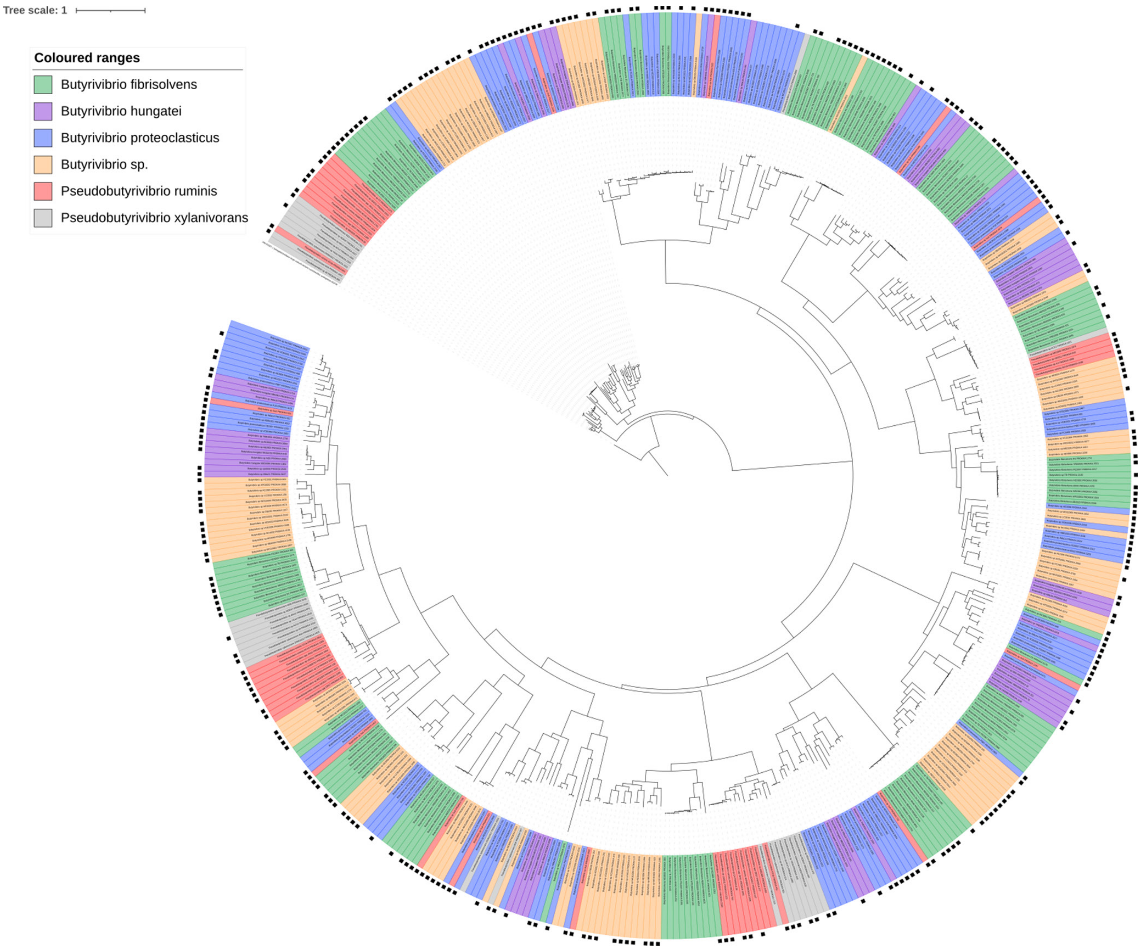

Tree scale: 1
Coloured ranges
🟩 Butyrivibrio fibrisolvens
🟪 Butyrivibrio hungatei
🟦 Butyrivibrio proteoclasticus
🟧 Butyrivibrio sp.
🟥 Pseudobutyrivibrio ruminis
⬜ Pseudobutyrivibrio xylanivorans

Fig. 7. Phylogenetic tree showing the relatedness of all GH family 2 genes found in all 71 strains used in this analysis. presence of a black square on the outermost layer indicates that the gene was found to be present in the Shi *et al.* metatranscriptome dataset [55]. The tree is rooted using a β-galactosidase large subunit sequence from *L. acidophilus* NCFM, which is coloured in black. Tree scale indicates number of substitutions per site.

Upon further checking of core and accessory gene assignments for the classical and ANI-defined *B. fibrisolvens* group for the 40 marker core orthologous genes used to reconstruct the 40 marker tree [11], it was noted that within the classically defined grouping that 11 of these were in the core and 22 in the accessory genes. Conversely, in the ANI-defined group, 26 were found within the core genes and none in the accessory genes. This again shows that the classical taxonomy is incorrect and provides further confidence in the fact that the members of the three ANI-defined groupings are correct, albeit they still possess open genomes. Functional annotation of the core and accessory genomes for the three groups also showed much similarity, with the only visual differences being that the proportion of unknown genes was higher in the accessory genome, and those related to translation, ribosomal structure and biogenesis were less numerous in the accessory genome (Fig. 5).

Gene orthology and paralogy

In order to evaluate gene ancestry and evolution, OrthAgogue was used to identify orthologous gene affiliations. Both genera share the majority of their orthologous genes, with 870 COGs being common to the two (Fig. 6). As a genus, *Pseudobutyrivibrio* has more common orthologous genes than *Butyrivibrio*, with 343 and 223 genes, respectively. *Pseudobutyrivibrio* has more unique COGs (595) than *Butyrivibrio* (251) (Fig. 6). The genus *Butyrivibrio* had the most inparalogous clusters (2596), and *Pseudobutyrivibrio* the fewest (259) (Table S2). Further analysis based on core/accessory designations on the correctly defined ANI taxonomy showed that most core genes were orthologues: 55, 60 and 26% for the *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups, respectively. Conversely, accessory orthologous genes constituted 36, 30 and 18% of the total genes for the *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups, respectively. Core genes contained 1, 1 and 12% inparalogues for the *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups, respectively. Conversely, accessory inparalogues constituted 1, 2 and 1% of the total genes for the *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups, respectively. Core genes also contained 3, 4 and 26% co-orthologues for the *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups, respectively. Conversely accessory co-orthologues

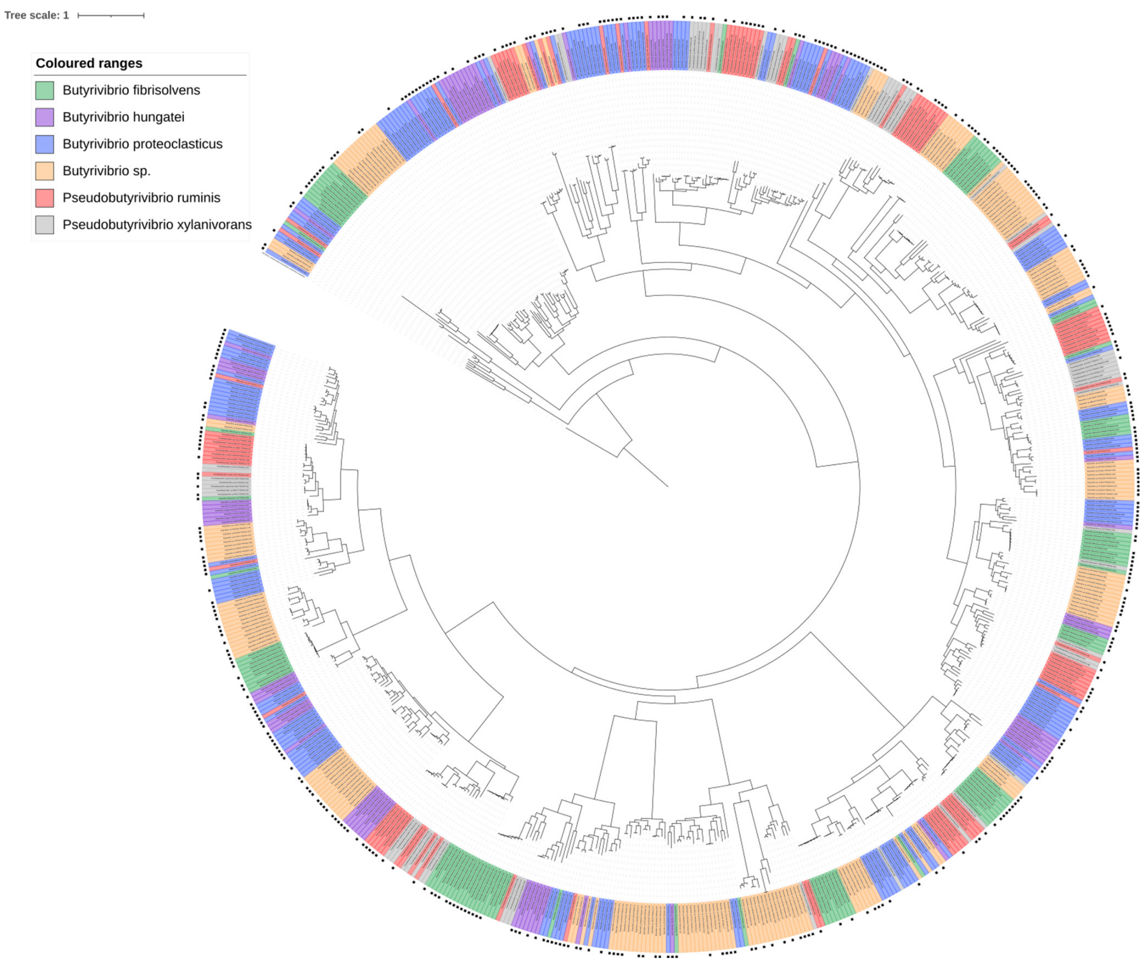


Fig. 8. Phylogenetic tree showing the relatedness of all GH family 3 genes found in all 71 strains used in this analysis. The presence of a black square on the outermost layer indicates that the gene was found to be present in the Shi *et al*. metatranscriptome dataset [55]. The tree is rooted using a *β*-*N*-acetylhexosaminidase sequence from *L. acidophilus* NCTC13720, which is coloured in black.

constituted 4, 3 and 17% of the total genes for the *B. fibrisolvens*, *B. hungatei* and *P. ruminis* groups, respectively.

GH haplotypes and evolution

Functional annotation of the GH families possessed by each strain showed a lot of similarity based on GH families and their abundances (Fig. S14). GH family phylogenetic trees show limited gene grouping, suggesting possession of a high degree of within-GH family enzyme isoform sequence diversity (Figs. 7–11). GH3 was the most abundant with 690 genes present, followed by GH13 with 681, GH43 with 543, GH2 with 463 and GH5 with 216 (Figs. 7–11, Supplementary Excel data 1). Pairwise comparisons of genes within these families showed extensive genomic diversity, with only 5.36% of GH2 genes sharing a pairwise percentage identity of greater than 50%. For GH3, this was 4.98 %, GH5 6.32 %, GH13 5.97 % and GH43 3.44% (Supplementary Excel data 2). Irrespective of relatedness, for each GH family, the majority of genes within were found in at least 1 of the 20 metatranscriptomes within the Shi *et al*. dataset [55] (Figs 7–11, Supplementary Excel data 3). For GH2, 67.17% of genes were found to be expressed, for GH3 62.87 %, GH5 56.28 %, GH13 66.57 % and GH43 59.41%. Of these, many had an RPKM >1, illustrating that the GH isoforms discovered are not an anomaly of the assembly and are actively expressed.

DISCUSSION

The role that *Butyrivibrio* and *Pseudobutyrivibrio* play in the rumen is not yet fully understood; however, they are known to be heavily involved in the metabolism of carbohydrates [23, 56], proteins [57] and lipids [58]. Indeed, they dedicate a large proportion of their genetic capacity to the breakdown and reassembly of complex polysaccharides, with the resulting simple sugars undergoing fermentation to produce butyrate, a major source of energy for the ruminant [16, 22, 56, 59–61]. In this study, we show that *Butyrivibrio* and *Pseudobutyrivibrio* are more genetically diverse than their classical taxonomy suggests. They also possess open pangenomes, as shown by low core gene proportions, and numerous gene haplotypes within multiple CAZyme families, which we hypothesize may provide metabolic

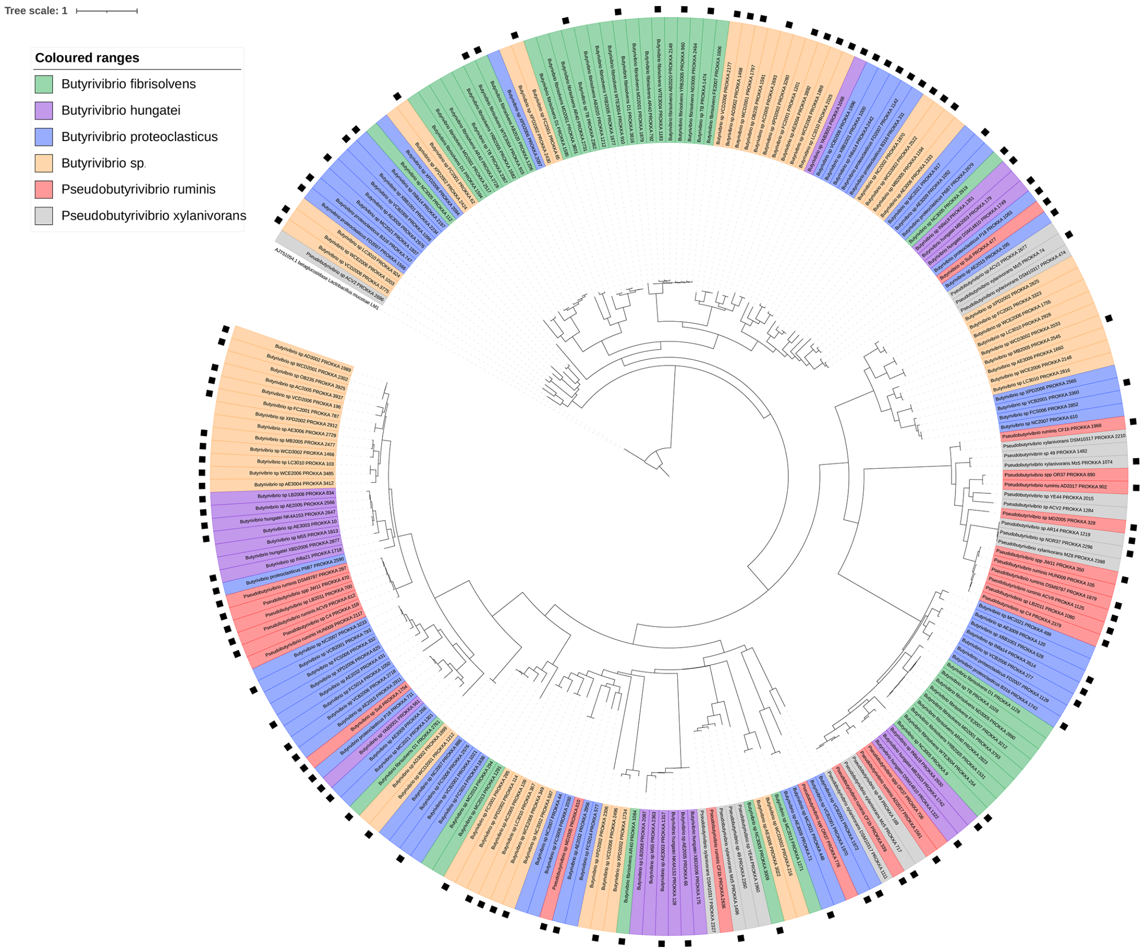

Tree scale: 1
Coloured ranges
🟩 Butyrivibrio fibrisolvens
🟪 Butyrivibrio hungatei
🟦 Butyrivibrio proteoclasticus
🟧 Butyrivibrio sp.
🟥 Pseudobutyrivibrio ruminis
⬜ Pseudobutyrivibrio xylanivorans

Fig. 9. Phylogenetic tree showing the relatedness of all GH family 5 genes found in all 71 strains used in this analysis. The presence of a black square on the outermost layer indicates that the gene was found to be present in the Shi *et al.* metatranscriptome dataset [55]. The tree is rooted using a *β*-glucosidase sequence from *Lactobacillus mucosae* LM1, which is coloured in black.

plasticity during dietary fluctuations. This study delves into the fundamental taxonomy, ecology and evolution of the *Butyrivibrio* and *Pseudobutyrivibrio* at a level not possible before the recent increase in available genomes [16].

A phylogenetic tree based on 40 conserved gene markers revealed groups that approximate to classical species, with the exception of *P. ruminis* strain CF1b. 16S rDNA phylogeny on fewer strains, performed by Kasperowicz *et al.* [62], showed that the CF1b strain groups closely with the type strain *P. ruminis* A12-1, which is concurrent with our own 16S rDNA findings. Whilst 16S rDNA analysis has historically been thought to be a reliable means of establishing distant relationships between organisms [63], extensive diversity has been found within the 16S rDNA of certain genomes [64, 65]. Furthermore, 16S rDNA has more recently been highlighted as providing poor resolution at a species level [9]. The 40 marker genes used are universal, single-copy genes that are highly conserved and appear to maintain a constant rate of horizontal transfer; as such, using these 40 markers is thought to provide a more resolved comparison [35]. Although both of these phylogenies broadly form the six classical taxa, some of these clades are not monophyletic, and as such this should not form the sole basis of their taxonomy. The ANI plot shows very little similarity overall between strains, yet several small groups with >95% ANI can be observed; the first of these is comprised of eight strains that were allocated to *B. fibrisolvens* in the previous phylogenies, the second being six strains of *P. ruminis* and the third four strains of *B. hungatei*. The alignment coverage plot similarly shows extensive dissimilarity, with few groups having coverage of >50%. Given that a 95% ANI cut-off value is commonly used to delineate species, and 50% coverage for genus [37], this suggests the presence of potentially 42 species in 32 genera. Wittouck *et al.* [66] used core nucleotide identity to reclassify the similarly structured *Lactobacillus* genus complex, stating that current bacterial taxonomy is not always consistent with bacterial evolutionary history, with some official taxa not being monophyletic (similar to the *Butyrivibrio* group). They proposed reclassification of 2459 genomes into 239 similar *de novo* species based on a 94% core nucleotide identity cut-off. Following this, another study re-evaluated the taxonomic structure of *Lactobacillus*, stating that micro-organisms that are

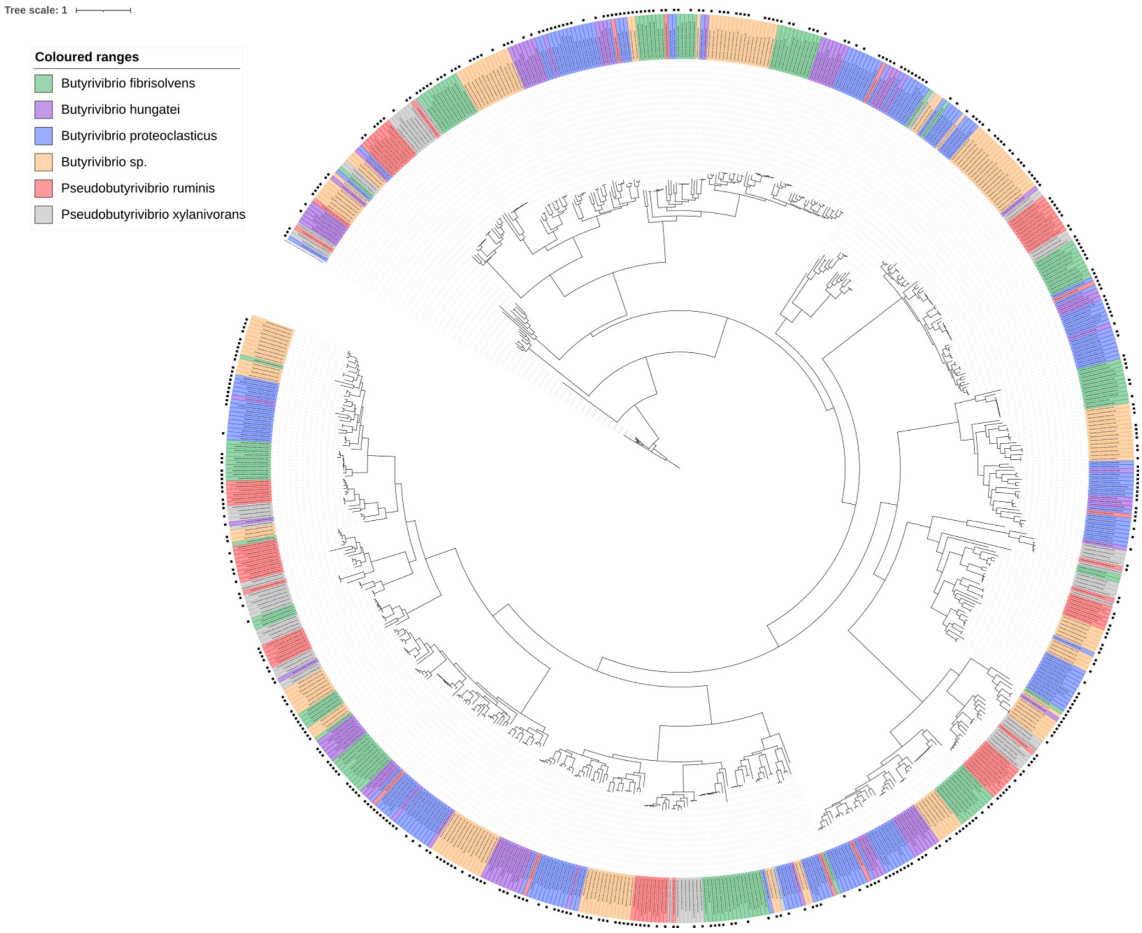

Tree scale: 1
Coloured ranges
Butyrivibrio fibrisolvens
Butyrivibrio hungatei
Butyrivibrio proteoclasticus
Butyrivibrio sp.
Pseudobutyrivibrio ruminis
Pseudobutyrivibrio xylanivorans

Fig. 10. Phylogenetic tree showing the relatedness of all GH family 13 genes found in all 71 strains used in this analysis. The presence of a black square on the outermost layer indicates that the gene was found to be present in the Shi *et al.* metatranscriptome dataset [55]. The tree is rooted using a sucrose phosphorylase sequence from *L. acidophilus* NCFM, which is coloured in black.

genetically distinct, as well as metabolically, ecologically and functionally diverse, were being grouped within the same genus, leading to high levels of genetic diversity. Using a combination of pangenomics, average amino acid identity (AAI), it was proposed that *Lactobacillus* be split into a further 24 genera, with the original emended genus bringing the total to 25 [67]. Given that ANI has been widely used for the delineation of species [9, 68, 69], we suggest that the genera *Butyrivibrio* and *Pseudobutyrivibrio* (and the species they currently encompass) be similarly split based on ANI into 42 species within 32 genera.

The increasing number of available bacterial genomes has allowed further research into microbial population genomics [70], which has revealed extensive intraspecific variability in prokaryotic genome content, and led to the study of pangenomes (all the gene families that have been found in the species as a whole), core genome ('essential' genes), and accessory genome ('dispensable' genes) [71]. This, alongside the analysis of orthologous genes (those derived from speciation events) and paralogous genes (those derived from duplication events), can give insight into the taxonomy and evolutionary divergence of a population. Pangenome analysis of our strains showed that, when analysed based on their classical taxon, *B. fibrisolvens* possesses the lowest percentage of core genes (2.45%), and *P. ruminis* the highest (10.38%), with both values being comparatively low, illustrating the low numbers of shared genes between strains. The previously acknowledged heterogeneity in *Butyrivibrio* [23, 24, 72] as it currently exists increases the probability of newly introduced genes being designated accessory. This leads to the artificial inflation of the accessory genome and the pangenome as a whole, if strains that are too dissimilar are allocated the same species, as may have often been the case given the historic tendency to classify new butyrate-producing bacteria as *B. fibrisolvens*, despite diversity and genetic relatedness [23, 73]. When based on the ANI re-classified groups, core genomes were larger, comprising means of 59.62, 68.50 and 65.23% for groups 1, 2 and 3, within *B. fibrisolvens*, *B. hungatei* and *P. ruminis*, respectively. Further accuracy checking of accessory and core designations was conducted by prospecting genes in both for the 40 biomarker genes used to define the 40 marker phylogenetic tree [11]. These biomarker genes should be core in all strains and be orthologous in nature; therefore, their presence in the accessory genes suggests over-inflation of the accessory genome

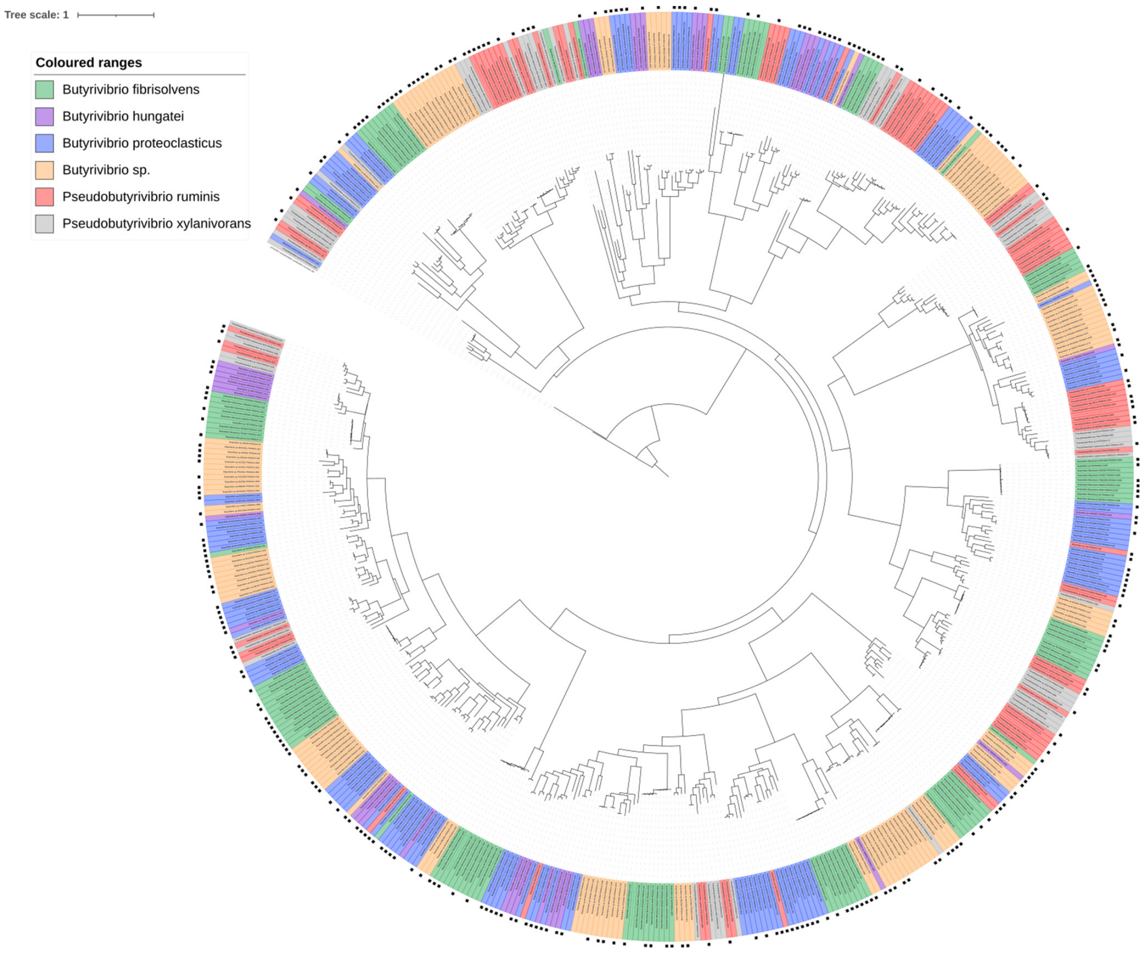

Tree scale: 1
Coloured ranges
🟩 Butyrivibrio fibrisolvens
🟪 Butyrivibrio hungatei
🟦 Butyrivibrio proteoclasticus
🟧 Butyrivibrio sp.
🟥 Pseudobutyrivibrio ruminis
⬜ Pseudobutyrivibrio xylanivorans

Fig. 11. Phylogenetic tree showing the relatedness of all GH family 43 genes found in all 71 strains used in this analysis. The presence of a black square on the outermost layer indicates that the gene was found to be present in the Shi *et al.* metatranscriptome dataset [55]. The tree is rooted using a β-xylosidase sequence from *Lactobacillus mucosae* LM1, which is coloured in black.

due to the presence of too much divergence in the strains. For the classical taxonomy, we found 22 of these genes within the classical taxonomy accessory genome of *B. fibrisolvens*, but none were found in the ANI-refined *B. fibrisolvens* group accessory genome. This suggests again that the classical taxonomy is incorrect and that the ANI re-definitions are highly likely to be more correct, although it should be noted that a reasonable diversity is still present in the ANI-defined groups. Indeed, these would still be considered to be open pangenomes when compared with, for example, *Pseudomonas aeruginosa*, whose core genes comprise 82–93% of their genomes using a 90% core definition [38]. Similarly, Lapierre and Gogarten [74] estimate that the typical bacterial genome would be composed of only 28% accessory genes.

When based on ANI re-defined taxonomy, functions relating to translation, ribosomal structure and biogenesis were found mainly in the core genome. This is likely due to the fact that a core genome is thought to comprise essential gene families, i.e. housekeeping genes [71], which are more likely to be present in a wider range of less-related bacteria. Likewise, it is perhaps unsurprising that the accessory group has more unknown genes, as these are likely to be very unique as they have yet to be characterized.

G+C mol% is consistently higher in the core genomes of each of the 71 strains when the pangenome is based on classical taxa, and the 17 strains in three groups formed by ANI. It has been suggested that G+C-rich genome segments can occur as a result of biased gene conversion following recombination, whereby DNA repair of mismatched bases holds a bias towards G+C nucleotides [75]. If this is assumed to be correct, this G+C bias in core genes could be explained by their retention over accessory genes, which are more readily lost and exchanged [76]; the longer these core genes are retained, the more they will be subjected to DNA repair, resulting in an increasing amount of G+C bases being incorporated into the core genome. Indeed, a recent study that analysed 731 prokaryotic strains from 36 species, 28 genera and 10 phyla found that the G+C mol% of the core genome was significantly higher than that of the whole genome, and that the G+C mol% of the accessory genome was significantly lower than that of the whole genome [77].

In terms of their evolution, the majority of COGs are shared across both the genera *Butyrivibrio* and *Pseudobutyrivibrio*, with 870 clusters being common to all of them. The genus *Pseudobutyrivibrio* group share a high proportion of COGs (343) and the *Butyrivibrio* group share slightly fewer at 223. When the number of genes found as part of a cluster of orthologues, inparalogues and co-orthologues are separated into core and accessory for the ANI-defined groups, the core genomes appear to be largely composed of orthologues as would be expected for related strains. It should be noted, however, that many publications exist which show that accessory genomes contain the most orthologues, which is likely due to the fact that their strains are so divergent that the partitioning of core and accessory genes is more difficult, as was seen for the analysis completed for the classical *B. fibrisolvens* grouping.

GHs are involved in the breakdown of carbohydrates, including many plant polymers, and are broken down into 111 families (http://www.cazy.org/) on the basis of amino acid similarity [78]. The rumen microbiome is exposed to strong diet-driven selection pressures, meaning that they must constantly compete for available sources of nutrition during dietary fluctuations [79]. The topology of the GH family trees is indicative of multi-gene families, i.e. groups of genes that have arisen from a common ancestor by duplication, and indicates the presence of a multitude of different isoforms. It is not uncommon for extensive gene sequence variation to be found within a bacterial family, with the resulting enzymes having different substrate specificities and yielding different products [80]. Ohta [81] further states that many multi-gene families are present in large numbers within a genome due to an increased demand for their gene product, with genes either retaining their function or diverging. Pairwise comparisons of genes representing the same GH family also show the vast diversity in the gene sequences and possibly suggest that there is too much diversity to draw phylogenetic trees. Irrespective, both approaches illustrate the fact that diversity in high in the families. Bacteria in functionally demanding environments are thought to possess a vast array of functional isomers, allowing resilience under dietary perturbations [82]. The fact that such a large proportion of the GHs were found within the Shi *et al.* dataset [55] confirms that they are actively expressed within the rumen and not artefacts of the genome assembly.

Conclusion

In conclusion, this study provides the most in-depth dataset on the phylogenetic systematics and evolution of the ruminal *Butyrivibrio* and *Pseudobutyrivibrio* to date. This study demonstrates remarkable genomic dissimilarity between strains that have previously been classified as the same species; this can be seen in the existence of outlier strains within the existing taxonomy in terms of phylogeny, G+C content, genome size and ANI, suggesting that they may be incorrectly classified. The genomes studied display very low per cent core genomes and high per cent accessory genomes when analysed in their current taxonomical groups. As such, we propose that their taxonomy be re-evaluated on the basis of their ANI to 42 species within 32 genera. Despite genomic variation, classical taxa appear to retain broadly similar high-level functional profiles, but possess a number of GH isoforms that we hypothesize facilitate metabolic plasticity and resilience under dietary perturbations.

Funding information

We acknowledge funding from the Knowledge Economy Skills Scholarships and the Department for the Economy who funded MPhil and PhD studies for S.E.P. We also acknowledge funding from BBSRC (BB/J0013/1; BBS/E/W/10964A-01) and RCUK Newton Institutional Link Funding (172629373).

Author contributions

S.A.H. and S.E.P., conceptualized the research and led the project. T.S., K.S.-T., F.G.S. and C.J.C., helped S.E.P. with the computational analyses and discussions regarding project direction. S.J.C., performed some computation analyses. S.E.P., S.A.H., C.J.C. and T.S., drafted the manuscript. All authors read and approved the final manuscript.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Bapteste E, Boucher Y. Epistemological impacts of horizontal gene transfer on classification in microbiology. *Methods Mol Biol* 2009;532:55–72.
2. Thane Papke R. A critique of prokaryotic species concepts. *Methods Mol Biol* 2009;532:379–395.
3. Schleifer KH. Classification of Bacteria and Archaea: past, present and future. *Syst Appl Microbiol* 2009;32:533–542.
4. Wayne L, Brenner D, Colwell R, Grimont P, Kandler O, *et al*. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Evol Microbiol* 1987;37:463–464.
5. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, *et al*. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019;10:5029.
6. Zhang W, Du P, Zheng H, Yu W, Wan L, *et al*. Whole-genome sequence comparison as a method for improving bacterial species definition. *J Gen Appl Microbiol* 2014;60:75–78.
7. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. *Nat Methods* 2013;10:881–884.
8. Rossi-Tamisier M, Benamar S, Raoult D, Fournier PE. Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. *Int J Syst Evol Microbiol* 2015;65:1929–1934.
9. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 2005;102:2567–2572.
10. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, *et al*. Toward automatic reconstruction of a highly resolved tree of life. *Science* 2006;311:1283–1287.
11. Cowan ST. Sense and nonsense in bacterial taxonomy. *J Gen Microbiol* 1971;67:1–8.
12. Goryunov DV, Nagaev BE, Nikolaev MY, Alexeevski AV, Troitsky AV. Moss phylogeny reconstruction using nucleotide pangenome of complete mitogenome sequences. *Biochemistry (Mosc)* 2015;80:1522–1527.
13. Moldovan MA, Gelfand MS. Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. *Front Microbiol* 2018;9:428.
14. Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* 2015;7:72–85.
15. Ricard G, McEwan NR, Dutilh BE, Jouany JP, Macheboeuf D, *et al*. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* 2006;7:22.

16. Seshadri R, Leahy SC, Attwood GT, Teh KH, Lambie SC, *et al*. Cultivation and sequencing of rumen microbiome members from the Hungate1000 collection. *Nat Biotechnol* 2018;36:359–367.

17. Henderson G, Cox F, Ganesh S, Jonker A, Young W, *et al*. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci Rep* 2015;5:14567.

18. Bryant MP. Bergey's manual of systematic bacteriology. Sneath P, Mair N, Sharpe H and Holt J (eds). In: *Bergey's Manual of Systematic Bacteriology vol. 2*. Baltimore: Williams and Wilkins; 1986. pp. 1376–1379.

19. Bryant MP, Small N. The anaerobic monotrichous butyric acid-producing curved rod-shaped bacteria of the rumen. *J Bacteriol* 1956;72:16–21.

20. Cheng KJ, Costerton JW. Ultrastructure of *Butyrivibrio fibrisolvens*: a gram-positive bacterium. *J Bacteriol* 1977;129:1506–1512.

21. Hespell RB, Wolf R, Bothast RJ. Fermentation of xylans by *Butyrivibrio fibrisolvens* and other ruminal bacteria. *Appl Environ Microbiol* 1987;53:2849–2853.

22. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, *et al*. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol* 2019;37:953–961.

23. Kopecny J, Zorec M, Mrazek J, Kobayashi Y, Marinsek-Logar R. *Butyrivibrio hungatei* sp. nov. and *Pseudobutyrivibrio xylanivorans* sp. nov., butyrate-producing bacteria from the rumen. *Int J Syst Evol Microbiol* 2003;53:201–209.

24. Moore W, Johnson J, Holdeman L. Emendation of *Bacteroidaceae* and *Butyrivibrio* and descriptions of *Desulfomonas* gen. nov. and ten new species in the genera *Desulfomonas*, *Butyrivibrio*, *Eubacterium*, *Clostridium*, and *Ruminococcus*. *Int J Syst Evol Microbiol* 1976;26:238–252.

25. Moon CD, Pacheco DM, Kelly WJ, Leahy SC, Li D, *et al*. Reclassification of *Clostridium proteoclasticum* as *Butyrivibrio proteoclasticus* comb. nov., a butyrate-producing ruminal bacterium. *Int J Syst Evol Microbiol* 2008;58:2041–2045.

26. Van Gylswyk N, Hippe H, Rainey F. *Pseudobutyrivibrio ruminis* gen. nov., sp. nov., a butyrate-producing bacterium from the rumen that closely resembles *Butyrivibrio fibrisolvens in ph*enotype. *Int J Syst Evol Microbiol* 1996;46:559–563.

27. Palevich N, Kelly WJ, Leahy SC, Denman S, Altermann E, *et al*. Comparative genomics of rumen *Butyrivibrio* spp. uncovers a continuum of polysaccharide-degrading capabilities. *Appl Environ Microbiol* 2020;86:e01993-19.

28. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.

29. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15:R46.

30. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

31. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016;44:W3–W10.

32. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, *et al*. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 2014;42:D633–D642.

33. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.

34. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 2008;9:R151.

35. Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS One* 2011;6:e22099.

36. Letunic I, Bork P. Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;44:W242–W245.

37. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 2016;8:12–24.

38. Ozer EA, Allen JP, Hauser AR. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa using* bioinformatic tools Spine and AGEnt. *BMC Genomics* 2014;15:737.

39. Harris CD, Torrance EL, Raymann K, Bobay LM. CoreCruncher: fast and robust construction of core genomes in large prokaryotic data sets. *Mol Biol Evol* 2021;38:727–734.

40. Zou W, Ye G, Zhang K, Yang H, Yang J. Analysis of the core genome and pangenome of *Clostridium butyricum*. *Genome* 2021;64:51–61.

41. Bentley S. Sequencing the species pan-genome. *Nat Rev Microbiol* 2009;7:258–259.

42. Lefebure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* 2007;8:R71.

43. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;11:472–477.

44. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 2017;34:2115–2122.

45. Ekseth OK, Kuiper M, Mironov V. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* 2014;30:734–736.

46. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;13:2178–2189.

47. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 2014;20:1983–1992.

48. Yin Y, Mao X, Yang J, Chen X, Mao F, *et al*. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2012;40:W445–W451.

49. Madeira F, Park YM, Lee J, Buso N, Gur T, *et al*. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019;47:W636–W641.

50. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;32:3047–3048.

51. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.

53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.

54. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;30:923–930.

55. Shi W, Moon CD, Leahy SC, Kang D, Froula J, *et al*. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res* 2014;24:1517–1525.

56. Kelly WJ, Leahy SC, Altermann E, Yeoman CJ, Dunne JC, *et al*. The glycobiome of the rumen bacterium Butyri*vibrio proteoclasticus B316*(T) highlights adaptation to a polysaccharide-rich environment. *PLoS One* 2010;5:e11942.

57. Cotta MA, Hespell RB. Proteolytic activity of the ruminal bacterium *Butyrivibrio fibrisolvens*. *Appl Environ Microbiol* 1986;52:51–58.

58. Paillard D, McKain N, Chaudhary LC, Walker ND, Pizette F, *et al*. Relation between phylogenetic position, lipid metabolism and butyrate production by different *Butyrivibrio*-like bacteria from the rumen. *Antonie van Leeuwenhoek* 2007;91:417–422.

59. Bond JJ, Dunne JC, Kwan FY, Li D, Zhang K, *et al*. Carbohydrate transporting membrane proteins of the rumen bacterium, *Butyrivibrio proteoclasticus*. *J Proteomics* 2012;75:3138–3144.
60. Marounek M, Petr O. Fermentation of glucose and xylose in ruminal strains of *Butyrivibrio fibrisolvens*. *Lett Appl Microbiol* 1995;21:272–276.
61. Palevich N, Kelly WJ, Leahy SC, Altermann E, Rakonjac J, *et al*. The complete genome sequence of the rumen bacterium *Butyrivibrio hungatei* MB2003. *Stand Genomic Sci* 2017;12:72.
62. Kasperowicz A, Stan-Glasek K, Guczynska W, Piknova M, Pristas P, *et al*. Sucrose phosphorylase of the rumen bacterium *Pseudobutyrivibrio ruminis* strain A. *J Appl Microbiol* 2009;107:812–820.
63. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, *et al*. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* 1985;82:6955–6959.
64. Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* 2004;186:2629–2635.
65. Beumer A, Robinson JB. A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Appl Environ Microbiol* 2005;71:8301–8304.
66. Wittouck S, Wuyts S, Meehan CJ, Noort van, Lebeer S. A genome-based species taxonomy of the *Lactobacillus* genus complex. *mSystems* 2019;4.
67. Zheng J, Wittouck S, Salvetti E, Franz C, Harris HMB, *et al*. A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae and Leuconostocaceae*. *Int J Syst Evol Microbiol* 2020;70:2782–2858.
68. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, *et al*. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018;68:461–466.
69. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
70. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, *et al*. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet* 2016;12:e1006280.
71. Gabaldon T, Koonin EV. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 2013;14:360–366.
72. Mannarelli B. Deoxyribonucleic acid relatedness among strains of the species *Butyrivibrio fibrisolvens*. *Int J Syst Evol Microbiol* 1988;38:340–347.
73. Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, *et al*. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet* 2015;11:e1004941.
74. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. *Trends Genet* 2009;25:107–110.
75. Huang S, Zhang S, Jiao N, Chen F. Comparative genomic and phylogenomic analyses reveal a conserved core genome shared by estuarine and oceanic cyanopodoviruses. *PLoS One* 2015;10:e0142962:11.:.
76. Hiratani I, Leskovar A, Gilbert DM. Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proc Natl Acad Sci U S A* 2004;101:16861–16866.
77. Bohlin J, Eldholm V, Pettersson JH, Brynildsrud O, Snipen L. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* 2017;18:151.
78. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 2014;42:D490–D495.
79. Ferrer M, Ghazi A, Beloqui A, Vieites JM, Lopez-Cortes N, *et al*. Functional metagenomics unveils a multifunctional glycosyl hydrolase from the family 43 catalysing the breakdown of plant polymers in the calf rumen. *PLoS One* 2012;7:e38134.
80. Mertz B, Gu X, Reilly PJ. Analysis of functional divergence within two structurally related glycoside hydrolase families. *Biopolymers* 2009;91:478–495.
81. Ohta T. Gene families: Multigene families and superfamilies. *eLS* 2006;1.
82. Rubino F, Carberry C, Waters SM, Kenny D, McCabe MS, *et al*. Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. *ISME J* 2017;11:932–944.