Phylogenetic systematics of *Butyrivibrio* and *Pseudobutyrivibrio* genomes illustrate possession of open genomes rich in orthologous accessory genes with an abundance of carbohydrate-active enzyme isoforms

Sharon Huws (s.huws@qub.ac.uk)
Queens university Belfast https://orcid.org/0000-0002-9284-2453
Timofey Skvortsov
Queen's University Belfast Faculty of Medicine Health and Life Sciences
Fernanda Godoys Santos
Queen's University Belfast School of Medicine Dentistry and Biomedical Sciences
Stephen Courtney
Queen's University Belfast School of Medicine Dentistry and Biomedical Sciences
Karen Siu Ting
Queen's University Belfast Faculty of Medicine Health and Life Sciences
Chris Creevey
Queen's University Belfast Faculty of Medicine Health and Life Sciences

Research

Keywords: rumen, Butyrivibrio, Pseudobutyrivibrio, pangenome, evolution, function, phylogeny, taxonomy

Posted Date: June 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34780/v1

Abstract

Background

Butyrivibrio and *Pseudobutyrivibrio* dominate in anaerobic gastrointestinal microbiomes, particularly the rumen, where they play a key role in harvesting energy from the diet. Within these genera, 5 rumen species have been classified (*B. fibrisolvens*, *B. hungatei*, *B. proteoclasticus*, *P. ruminis* and *P. xylanivorans*); nonetheless, the phylogeny and evolution of these genera is still unclear. Given the recent increase in available genomes, a re-investigation of the phylogenetic systematics and evolution of *Butyrivibrio* and *Pseudobutyrivibrio* is timely.

Results

We show, using both a 16S rDNA and 40 gene marker phylogenetic tree, that 6 species, namely 1. *P. ruminis*, 2. *P. xylanivorans*, 3. *B. fibrisolvens,* 4. *Butyrivibrio sp.*, 5. *B. hungatei*, and 6. *B. proteclasticus* likely exist. Pangenome analysis at 100% core definition showed a high abundance of accessory genes (91.50 to 99.34%) compared with core genes (0.66 to 8.50%), illustrating possession of very open genomes. Across the 71 genomes, 870 COGs (clusters of orthologous genes) were shared by all taxa, suggesting evolution through speciation from a common ancestor. Further analysis of Carbohydrate-Active Enzymes (CAZymes) genes show that most are within the accessory genome and orthologous in descent with numerous within-family CAZyme isoforms apparent, CAZyme family tree lineages show that these isoforms largely group according to the 6 species, suggesting extensive horizontal gene transfer within these families.

Conclusions

We show the extensive genomic variation found within *Butyrivibrio*, and to a lesser extent, *Pseudobutyrivibrio*. and demonstrate the existence of a new *Butyrivibrio* species. The *Butyrivibrio* and *Pseudobutyrivibrio* genomes are very open with very low % core genomes and high % accessory genomes., and possess a number of GH isoforms that we hypothesise facilitate metabolic plasticity and resilience under dietary perturbations. This study utilizes all currently available genomes and consequently provides a major advancement in our understanding of these important anaerobic bacteria.

Background

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]. Originally, in the late 19th century, 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 DNA-DNA hybridisation in the mid-late 20th century. More recently, we have also used genotypic analyses, multilocus sequence analyses, average nucleotide identity, whole genome analyses etc [3,4]. 16S rDNA became a popular metric in the 1980’s, with recommendations made that organisms sharing greater than 98% 16S rDNA should be classified as a single species [5]. This was further developed to whole genome alignments [6] and phylogenetic clustering [7], nonetheless both of these molecular-based tools faced scrutiny for their seemingly arbitrary cutoff values [8]. 16S rDNA was also criticised on the basis that only a single gene is used as a point of comparison [9]. As a result of the extensive variation in bacterial classification, 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 [3, 10]. More recently, pangenomic analyses have also been suggested as potential methods for defining prokaryotic species [11]. For example, Moldovan and Gelfand (2018; 12) propose a new procedure for the definition of bacterial species that combines phylogeny and pangenomes. In their model, they use both a strict species definition and a weak definition. The strict definition states that a species should be monophyletic in a sequence-based tree and should be composed of a genetically similar strain set, and should contain the maximum amount of strains that satisfy both of the previous criteria. The weak definition allows the species to be either monophyletic (with paraphyletic clades potentially being sub-species) or polyphyletic. Furthermore, Goryunov *et al*. (2015, 13) use a nucleotide pangenome (that is, a set of aligned nucleotide sequences of orthologous gene fragments covering all genomes) to create phylogenies of mosses, highlighting the potential usefulness of pangenomes in taxonomy once more. Likewise, orthologous gene analysis has been used for phylogenetic purposes when taxonomic discrimination is challenging [14], whilst also providing information on gene evolution and therefore strain evolution.

The rumen microbiome is an example of a taxonomically ambiguous environment, with horizontal gene transfer being rife due to the contained nature of the rumen, and 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, most of which are bacterial. More recent metataxonomic studies,

including the global rumen census, show that a core rumen microbiome can be found in various ruminant breeds across varying geographical locations [17]. These studies show that *Prevotella*, *Butyrivibrio*, and *Ruminococcus*, as well as unclassified *Lachnospiraceae*, *Ruminococcaceae*, *Bacteroidales*, and *Clostridiales* make up the "core bacterial microbiome" across breeds and geographical locations [17]. Nonetheless, the taxonomy of *Butyrivibrio* and *Pseudobutyrivibrio* remains a topic of debate.

Butyrivibrio were first described in 1956 and using classical morphological and biochemical taxonomy were described as being motile, Gram negative, slightly curved rods that produced large amounts of butyric acid via glucose fermentation [18,19]. Despite this initial description, they were later found to be Gram positive, possessing derivatives of teichoic acid in their cell membrane [20]. It was noted upon discovery that there was extensive variation within the genus, and suggested that this variation may lead to difficulties in defining species-specific patterns, despite butyrate production being a commonality [18, 19, 21]. The first isolates were called *Butyrivibrio fibrisolvens*, named after their importance in the digestion of fibrous constituents in ruminant feed [18,19], enabled by their vast array of carbohydrate active enzymes [22]. From here, newly isolated strains were routinely classified as *B. fibrisolvens* despite morphological variation and vast genetic diversity [23] until 1976, when *B. crossotus,* a predominantly human isolate (also being found in the rumen on occasion in low abundance), was first described [24].

Currently there are 4 *Butyrivibrio* species which reside in the rumen: namely, *B. fibrisolvens*, *B. crossotus, B. hungatei* and *B. proteoclasticus* (Fig. 1) [18, 24, 25]. In 2008, *B. proteoclasticus* was reclassified, originally being *Clostridium proteoclasticum*, on the basis of phylogenetic placement, DNA GC content, and physiological traits [25]. In 1996, a non-succinate-fermenting bacterium, was isolated that closely resembled *B. fibrisolvens* but was found to vary sufficiently based on 16S rDNA, GC content and cellular fatty acid content, which was named *Pseudobutyrivibrio ruminis* [26]. Later on in 2003 another species of *Pseudobutyrivibrio* was taxonomically classified as *P. xylanivorans* based upon fermentation characteristics, DNA GC content and 16S rDNA dissimilarity to *Butyrivibrio* spp. [23]. Despite these additional species being named, the *Butyrivibrio* and *Pseudobutyrivibrio* still possess untapped phylogenetically and genetic diversity [27]. The challenges of defining prokaryotic taxonomy in general are also compounded by the diverse approaches applied to bacterial systematics [12].

The aims of this study were to re-investigate phylogeny, gene-level functional divergence and evolution in predominantly ruminal *Butyrivibrio* and *Pseudobutyrivibrio* using all publically available genomes. This study also investigated gene-centric evolution in the ruminal *Butyrivibrio* and *Pseudobutyrivibrio* in relation to their pangenomes. Many of the bacterial genomes in the Hungate collection are from the genus *Butyrivibrio* and *Pseudobutyrivibrio*, which makes this study timely, and enables a paradigm shift in our fundamental understanding of these genera.

Results

Functional similarity

Functional analysis was completed for all 71 ruminal *Butyrivibrio* and *Pseudobutyrivibrio* genomes used in this study using EggNOG-mapper (Table 1; Fig. 2). These data show very little variability at high-level functional categories between the genomes, and highlight that all the strains have a large proportion of 'unknown' genes (i.e. those not annotated by EggNOG-mapper). Overall, strains have predominant functions relating to carbohydrate transport and metabolism (average 8.91%), cell wall, membrane and envelope genesis (average 8.05%), and amino acid transport and metabolism (average 7.24%). The percentage of these functions stay fairly constant across all strains (standard deviation 0.80%, 0.78% and 0.50% respectively), although some exceptions to this can be seen. Specifically, strain YAB3001 has 188 genes of 3095 (6.07% of total) annotated as being part of an amino acid transport and metabolism pathway, compared with strain NOR37 which has 183 of 2184 (8.40% of total) attributed to the amino acid transport and metabolism pathway. The average percentage for this category across all strains is 7.24%. The general consistency in high-level gene function can also be seen when the functional categories are compared at a genus level (Supplementary Fig. 1).

Gene and Genome-level similarity

Despite the broad similarity seen in high-level functional categories (Fig. 2 and Supplementary Fig. 1), there may be substantial differences at the gene level. In order to evaluate gene level similarities for the 71 strains, phylogenetic trees based on 16S rDNA and 40 conserved marker genes were first constructed to examine phylogenetic relatedness (Supplementary Fig. 2, Fig. 3 respectively). The 40 marker phylogeny revealed six species/clades, containing: 1. *P. ruminis*, 2. *P. xylanivorans*, 3. *B. fibrisolvens,* 4. *Butyrivibrio sp.*, and 5. *B. hungatei*, and 6. *B. proteoclasticus,* although it should be noted that *P. ruminis* and *P. xylanivorans* are phylogenetically very close, as are *B. hungatei*, and *B. proteoclasticus* (Fig. 3). Similar patterns were seen in the 16S rDNA based phylogenetic tree, although the clustering

was less distinct in some instances (Supplementary Fig. 2). Similar groupings can be seen again in the 3D scatter plot (Fig. 4), with *P. ruminis* and *P. xylanivorans* clustering very closely as well as *B. hungatei* and *B. proteoclasticus*.

In order to compare the strains on a genome level Average Nucleotide Identity (ANI) was calculated using PyANI [28]. These data show that the vast majority of strains have less than 75% nucleotide identity (Supplementary Fig. 3). Several small clusters can be seen with nucleotide identity greater than 75%. Groups formed by PyANI are somewhat similar to those seen in the 40 marker tree (Fig. 3), although anomalies can be seen. When dendrograms based on genome coverage are plotted, the data show that many of the strains have <50% coverage, meaning that <50% of the genome can be compared due to a high level of dissimilarity (Supplementary Fig. 4). Due to this reason the strength of the comparative data is weak when coverage is <50%, which in itself shows the level of dissimilarity across strains.

These data as a whole are conflicting in terms of enhancing the taxonomic assignment of the ruminal *Pseudobutyrivibrio* and *Butyrivibrio*. Nonetheless, the 40 marker tree allowed resolution of all 5 previously classified *Butyrivibrio* and *Pseudobutyrivibro* spp. to varying extents, whilst identifying another possible *Butyrivibrio* clade/species. The similarity between the 40 marker data and the 16S rDNA and 3D GC content/Genome Size/Number genes made it a logical choice on which to base further analysis.

The 40 marker phylogenetic tree suggests that *Butyrivibrio* spp. NC3005, MC2013 and TB are actually *B. fibrisolvens*; *Butyrivibrio* spp. INlla14, XBB1001, VCB2006, AE3009, MC2021, FCS006, NC2007, VCB2001, XPD2006, AE2032, FCS014, AE2015 and Su6 are actually *B. proteoclasticus*; *Butyrivibrio* spp. INlla18, INlla21, NK4153, AE2005, AE3003, M55, LB2008 and YAB3001 are actually *B. hungatei*; *Pseudobutyrivibrio* spp. MD2005, OR37, LB2011, JW11 and C4 are actually *P. ruminis*; *Pseudobutyrivibrio* spp. AR14, NOR37, YE44, ACV2 and 49 are actually *P. xylanivorans*; *Butyrivibrio* spp. NC2002, WCD3002, MB2005, AE3006, VCD2006, WCD2001, OB235, AD3002, AC2005, XPD2022, FC2001, AE3004, LC3010 and WCE2006 actually constitute a potentially new species. As such, all subsequent analyses were completed based on these 6 species/clades.

Pangenomics

In order to further scrutinise this taxonomic ambiguity, the pangenomes were investigated using Spine software [29] based on the 6 species/clades identified. A range of Spine cut-off parameters (% accessory definition, the number of genomes a gene that has to be found within to be considered core and % nucleotide identity) were assessed to ensure optimal non-biased parameters were chosen for downstream analysis (Supplementary Fig. 5 and 6). These data show that, as expected, the higher the core definition percentage is, the greater the proportion of accessory genes. There appears to be a sharp increase in the percentage of the genome that is defined as accessory when the core definition setting is increased from 60 to 70% in *P. ruminis*, and 70 to 90% in *B. fibrisolvens*. All other groups show a more consistent estimate of the accessory genome regardless of the core definition percentage used (Supplementary Fig. 5). Contrastingly, as the gene nucleotide identity percentage cutoff increases, there is little change in the proportion of accessory genes observed (Supplementary Fig. 6). It is challenging to define the 'correct' cut-off parameters for pangenome analysis, which ultimately correspond to the biological situation. For downstream analysis the parameter of 100% core definition was used, i.e the gene has to be present in all genomes (as per 30), as well as the default 85% gene nucleotide identity cut-off. The total number of core genes per clade decreases, as does the core/pangenome ration, with increasing stringency (Table 2).

At a 100% core definition, *B. fibrisolvens* has an average core GC content of 44.94%, and an accessory GC content of 40.57%. *B. hungatei* has 44.73% and 41.04% for average core and accessory GC contents respectively, *B. proteoclasticus* 45.23% and 42.70%, *Butyrivibrio sp.* 45.14% and 41.58%, *P. ruminis* 43.39% and 39.06%, and *P. xylanivorans* 43.69% and 39.12%. *Pseudobutyrivibrio* strains had a greater difference than *Butyrivibrio* strains, with an average difference of 4.32% compared to 3.38%. The greatest difference was seen in strains of *P. xylanivorans*, with an average 4.57% difference in GC content between core and accessory genome, and the least difference is in *B. proteoclasticus* with 2.52%. Core genes across each taxon appear to have a higher GC% than their respective accessory genomes (with an average of 44.57% in the core genome, and 40.95% in the accessory at a core definition of 90%) (Supplementary Table 1).

Functional annotation of species/clades, when split into core and accessory genomes (Fig. 5A and B) shows greater functional diversity within the accessory genome. Invariably, a large proportion of the core genome appears to be dedicated to translation, ribosomal structure, and biogenesis. *B. fibrisolvens* appears to dedicate the largest proportion of its core genome to this category. This can be seen to a much lesser extent in the accessory genomes, with no particular functional role dominating to the same extent. Functional categories appear to be relatively evenly distributed throughout all six accessory genomes. The most evident outlier in the core genome is the *B. fibrisolvens* core, which has 223 genes in the "Translation, ribosomal structure and biogenesis" category, which is 70.35% of all its core genes. The average for this category in the core genome is 53.18%. The largest category is that of proteins with unknown function, which is much more prevalent in the accessory genome of each taxon. The average composition of unknown genes in the core genomes is 1.70%, whilst for the accessory genome this is 24.65%. *B. fibrisolvens* seems to have the least diverse core genome, having genes from only 8 of the 21

functional categories, followed by *B. proteoclasticus* with 9. *B. fibrisolvens* and *B. proteoclasticus* also have the fewest core genes annotated, with 317 and 502 respectively. *B. hungatei* has 989, *Butyrivibrio sp.* 1445, *P. xylanivorans* 1541, and *P. ruminis* 1635.

ClustAGE plots show high levels of genomic dissimilarity within species level groups in terms of Accessory Genomic Elements (AGEs), with each concentric ring representing a genome in the clade (shown by the key), and the AGE positioning in the circle representing their size (Supplementary Fig. 7-12). The plot for *B. fibrisolvens* 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 four genomes out of 11 (AB2020, MC2013, NC3005, and D1) possessing an AGE here (Supplementary Fig. 7). In all species level groups, the most dissimilarity seems to be in smaller gene fragments (Supplementary Fig. 7-12). It should be noted that the minimum AGE size represented in each plot is 1500 bp.

Evolution of the *Butyrivibrio* and *Pseudobutyrivibrio* genera

In order to evaluate gene ancestry and evolution OrthAgogue [31] was used to identify orthologous gene affiliations. Orthologous gene distribution on a clade/species level shows that the vast majority of orthologous gene clusters are shared by all species level groups (870) (Fig. 6). As a genus, *Pseudobutyrivibrio* has more common orthologous genes than *Butyrivibrio*, with 343 and 223 genes respectively. *B. fibrisolvens* has the most unique orthologous genes with 143, followed by *P. xylanivorans* with 132, *B. hungatei* with 121, *Butyrivibrio sp.* with 90, *P. ruminis* with 27 and *B. proteoclasticus* with 17 (Fig. 6). Comparison on a genus level clearly shows that both genera share the majority of their orthologous genes, with 870 clusters of orthologous genes (COGs) being common to the two. *Pseudobutyrivibrio* has more unique COGs, with 595 whilst *Butyrivibrio* has 251 unique COGs (Supplementary Fig. 13). Out of all taxa, *Butyrivibrio sp.* shares the most amount of COGs across all its members with 1771, followed by *P. xylanivorans* with 1674, *B. hungatei* with 1602, *P. ruminis* with 1585, *B. proteoclasticus* with 1562 and *B. fibrisolvens* with 1502 (Supplementary Fig 14-19). At taxon level, *B. proteoclasticus* has the most inparalogous clusters with 920. This is followed by *Butyrivibrio sp.* (778), *B. fibrisolvens* (460), *B. hungatei* (438), *P. ruminis* (305), and *P. xylanivorans* (259) (Supplementary Table 2). No inparalogous genes were shared across all species. *B. proteoclasticus* had the genome with the most inparalogs, with strain FCS014 having 178. *B. fibrisolvens* had the fewest total inparalogous genes, with strain D1 having 53, the most of any member of that clade/species (Supplementary Table 2). Further analysis showed that the majority of accessory genes were orthologs; 892 orthologs, 45 inparalogs and 80 co-orthologs, being 87.71%, 4.42%, and 7.87% respectively of a total of 1017 accessory homologous genes. Conversely the majority of core genes were annotated as inparalogous in terms of their evolution; 10 orthologs and 240 inparalogs, with 0 co-orthologs. This equates to 4.0% and 96.0% respectively of a total of 250 core homologous genes. Combined, this gives an overall total of 1269 genes. Of these, 10 are core orthologs (0.79%), 242 are core inparalogs (19.07%), 45 are accessory inparalogs (3.55%), 80 are accessory co-orthologs (6.30%), and 892 are accessory orthologs (70.29%).

Glycosyl hydrolase 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. 7; Supplementary Fig. 20-25). However, the GH family-level phylogenetic trees show limited gene clustering and GHs from the same family and within the same strain have a tendency to cluster together, as can be seen on pruned versions of the trees (Supplementary Fig. 26 and 27). These phylogenetic trees therefore show that the genera *Butrivibrio* and *Pseudobutyrivibrio* possess a high degree of within GH family enzyme isoforms. Irrespective, GH3 was the most abundant with 690 genes present, followed by GH13 with 681, GH43 with 543, GH2 with 463, and GH5 with 216. Homolog types across the GH2 genes show very few inparalogs with most genes being co-orthologous or orthologous across all species/clades (Fig. 8). Evolutionary distribution across GH3 is fairly similar (Fig. 9), although there is a greater proportion of inparalogous genes, which also appear to be more evenly distributed across the species/clades. Only 6 inparalogs have been annotated in the family GH5 (Fig. 10), and 34 co-orthologs, which is comparatively few compared with the higher number of orthologs. The co-orthologs form two broad clusters, and are only found in the *Butyrivibrio sp.* and *B. proteoclasticus* clades. There are no co-orthologous genes found in *B. fibrisolvens*, and very few in *B. proteoclasticus*. GH13, interestingly, has more inparalogs than co-orthologs, with 70 and 48 respectively (Fig. 11). There is overlap here, with many of these genes being annotated as orthologs, inparalogs and co-orthologs simultaneously which is a function of the algorithms used by Orthagogue. GH43 has a high proportion of co-orthologs, and fewer inparalogs (Fig. 12). The co-orthologs appear to be distributed across the entire tree, but with a higher density towards the far end of the tree. For each of the five GH families, the majority of genes were found in one of the 20 metatranscriptome datasets within the Shi *et al.* (2014) dataset (Fig 8-12; Suppl. Excel 1). 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 (Reads Per Kilobase of transcript, per Million mapped reads) value of over 1. These data illustrated 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 a wide variety of carbohydrates [23, 32] proteins [33], and lipids [34]. Indeed, *Butyrivibrio* and *Pseudobutyrivibrio* 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, 32, 35]. In this study we show that *Butyrivibrio* and *Pseudobutyrivibrio* are genetically highly diverse, but can be classified within six species/clades, namely *B. fibrisolvens*, *B. hungatei*, *B. proteoclasticus*, *Butyrivibrio* sp., *P. rumins* and *P. xylanivorans*. These genera also contain very open genomes with a major proportion of the genes being part of the accessory genome as opposed to the core genome, as shown by the low core/pangenome ratio values. We also show that these bacterial genera possess a diversity of CAZymes and numerous gene haplotypes within each CAZyme family which we hypothesise may provide metabolic 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].

Based on strain level comparisons, functional categories appear to be relatively consistent across all genomes. It should be of no surprise that a large proportion of each genome (an average of 23.94% across each genome) has been annotated as “Function unknown”, given that it is not uncommon for genomes to contain genes of unknown function, and our understanding of gene function tends to be lacking leading to poor annotations. For example, when the “minimal bacterial genome” of *Mycoplasma mycoides* was produced (that is, the smallest set of genes within that species that is capable of supporting life), 23.95% of that genome was composed of genes with an unknown function [36]. *Butyrivibrio* and *Pseudobutyrivibrio* are well known carbohydrate-degrading bacteria [23, 26, 37, 38], due to their expansive CAZyme repertoire. Therefore, it is unsurprising that the “Carbohydrate transport and metabolism” functional SEED category is the most enumerate in the strains of *Butyrivibrio* and *Pseudobutyrivibrio* studied. “Cell wall/membrane/envelope biogenesis” and “Amino acid transport and metabolism” categories were also highly prevalent, likely due to the fact that they contain many housekeeping genes. Given that these strains all share the same environment, it is logical that they fulfil the same broadly functional tasks, and therefore have similar functional annotations.

A phylogenetic tree based on 40 conserved gene markers [39, 40] revealed groups that approximate to classical species, with the exception of *P. ruminis* strain CF1b, and another species/clade of *Butyrivibrio* was also evident. 16S rDNA phylogeny on fewer strains performed by Kasperowicz *et al.* (2009; 41) 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 is thought to be a reliable means of establishing distant relationships between organisms due to it’s high information content, conserved nature, and universal distribution [42], high levels of diversity have been found within the 16S rDNA genes of certain genomes [43, 44]. 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 a result of this, using these 40 markers are thought to provide a more resolved comparison [40]. Although we conclude that *Butyrivibrio* and *Pseudobutyrivibrio* can be classified within six species/clades, namely *B. fibrisolvens*, *B. hungatei*, *B. proteoclasticus*, *Butyrivibrio* sp. (a potentially new species/clade), *P. ruminis*, and *P. xylanivorans*, we note that *P. ruminis* and *P. xylanivorans* are phylogenetically very close as are *B. hungatei* and *B. proteoclasticus*. *B. hungatei* forms a paraphyletic clade, and it could therefore be suggested that it is a sub-species of *B. proteoclasticus* [12].

The increasing number of available bacterial genomes has allowed further research into microbial population genomics [45], which has revealed extensive intraspecific variability in prokaryotic genome content, and led to the coining of terms such as “pangenome” (all the gene families that have been found in the species as a whole), “core genome” (‘essential’ gene families that are found in all members sequenced thus far), and “accessory genome” (‘dispensable’ genes that are not in every genome) [46]. This, alongside the analysis of orthologous genes (those derived from speciation events) and paralogous genes (those derived from duplication events) can give an in-depth insight into the taxonomy and evolutionary divergence of a population. Gabaldón and Koonin (2013; 47) stated that orthology is the most accurate way of describing differences and similarities in the composition of genomes from different species because orthologs by definition trace back to an ancestral gene that was present in a common ancestor of the compared species. Pangenome analysis of our strains showed that *B. fibrisolvens* as a species/clade possesses the lowest percentage of core genes (2.45%), and *P. ruminis* the highest (10.38%), with both values being comparatively low illustrating that the genera have ‘open’ pangenome, meaning that as additional strains are introduced, the pangenome increases in size [29].

The high level of genetic variation already known in *Butyrivibrio* (23, 24, 48) increases the probability of newly introduced genes being accessory. It is also likely that *B. fibrisolvens* has the highest genotypic diversity, given the historic tendency to classify newly discovered strains of *Butyrivibrio* as *B. fibrisolvens* on the basis of common phenotypic and metabolic characteristics, despite their vast diversity and genetic relatedness [23, 49]. Core genomes, when viewed on a species/clade basis, share a high proportion of genes involved in translation, ribosomal structure, and biogenesis. This is likely due to the fact that a core genome is thought to comprise essential gene families, i.e. housekeeping genes, whilst the accessory genome is more likely to encode genes that confer functional variation [46].

GC% is consistently higher in the core genomes of each of the 71 strains. Whilst intra-genomic heterogeneity is common [50], it does not explain the evident divide between core and accessory genes. It has been suggested that GC rich segments can occur as a result of biased gene conversion (BCG) following recombination, whereby DNA repair of mismatched bases holds a bias towards GC nucleotides [51]. If this is assumed to be correct, this GC bias in core genes could be explained by their retention over accessory genes, which are more readily lost and exchanged, and are generally less conserved [52]. The longer these core genes are retained, the more they will be subjected to DNA repair, resulting in an increasing amount of GC bases being incorporated into the core genome. The BCG model is, however, contentious, with GC content being thought to play a role in several cellular processes such as replication and expression regulation [53]. Similarly, it has been suggested that there may be a universal mutation bias towards AT nucleotides as a result of single nucleotide polymorphisms (SNPs), suggesting that there may be pressure to retain high GC content [54. 55] state that there is a bias towards genes that are highly expressed having a higher GC content, indicating that core genes may be more highly expressed than accessory genes.

The vast majority of Clusters Of Orthologous Genes (COGs) are shared across all six species/clades, with 870 clusters being common for all of them. Both *Pseudobutyrvibrio* species/clades share a high proportion of clusters, with 343 COGs being shared between the two. The four *Butyrivibrio* groups share slightly fewer at 223, and *B. fibrisolvens* appears to have the most clade-unique COGs with 143. This could, again, be attributed to intraspecies genomic diversity. When the number of orthologs, inparalogs and co-orthologs are separated into core and accessory, the core genomes appear to be largely composed of inparalogs (242), only 10 orthologs, and no co-orthologs. Conversely, 45 inparalogs, 892 orthologs, and 80 co-orthologs were annotated as being in accessory genomes. It is possible that the core genome, which is thought to contain genes that are essential for life, has such a high number of inparalogs due to a high transcription demand. Generally, organisms with multi-copy genes yield higher expression levels of these proteins [56]. This is supported by Hutchison *et al*. (2016, 36), who state that it is common for bacteria to possess multiple copies of genes that are involved in essential, or quasi-essential, functions, and that these genes may or may not be paralogs. Despite this, it is thought that there is a positive selection towards deletion of superfluous genes given that a smaller genome facilitates faster replication times [57].

The high proportion of orthologs in the accessory genome could potentially be explained by the presence of multigene families (MGF) within the accessory genome. Walsh and Stephan (2008, 58) state that gene families are assumed to be derived from a common ancestor, meaning that they could be classified as orthologous. This could be linked to the tendency for microorganisms in complex environments possessing multiple isomers of the same gene. This confers not only competitive advantage, but also a level of environmental robustness that facilitates metabolism of a wide variety of substrates in a changing environment [59]. It is also logical that many orthologs will be found in the accessory genome, as they’re more likely to be single copy within a strain (given that they do not by definition undergo within-genome duplication, unlike inparalogs), and are therefore more likely to be lost from a genome than inparalogs. In being annotated as an ortholog, a given gene is unlikely to have any recent within-genome duplications (or else it would be annotated as an inparalog), and yet in being core it must be present in 100% of genomes. These criteria, when applied simultaneously, are restrictive. Dagan and Martin (2006; 60) state that core orthologous genes should account for approximately one percent of genes within a bacterial population. Given the already small collection of core genes found in *Butyrivibrio* and *Pseudobutyrivibrio*, it should not be surprising that these genes account for slightly less than this, at 0.79%, although this is confounded by the fact that many genes were annotated as being both orthologous and inparalogous.

Glycosyl hydrolases (GH) are involved in the breakdown of carbohydrates, including many plant polymers, and are broken down into 111 families (61; http://www.cazy.org/) on the basis of amino acid similarity [62]. Given that ruminants are well known for their ability to degrade diverse plant polymers at high rates, it should not be a surprise that they possess a vast array of GH enzymes. 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 [63]. The clustering that can be seen in all GH families whereby clusters of enzymes within the same species/clade, form a similar clustering to the 40 marker trees, is typical of a multigene family, i.e. a group of genes that have arisen from a common ancestor by duplication. These genes therefore have similar functions and a high level of sequence similarity [64]. This is supported by the annotation of many of the GHs as orthologs. Strain *P. ruminis* CF1b is again the exception here, grouping with *P. xylanivorans* more closely, suggesting that it may actually belong with the *P. xylanivorans* species/clade. Although species/clade level clustering can be seen, a wide variety of GH isoforms is evident. It is not uncommon for extensive sequence variation to be found within a bacterial family, with the resulting enzymes having different substrate specificities and yielding different products [65]. Ohta (2008; 66) further state that many multigene families are present in large numbers within a genome due to an increased demand for their gene product, with genes either being clustered or dispersed throughout the genome. The clustered genes may retain overlapping functions whilst the dispersed genes may diverge to a greater extent. This may again be implicated in the tendency for bacteria in functionally demanding environments, such as the rumen, to possess a vast array of functional isomers allowing resilience under dietary perturbations

[59]. The fact that such a large proportion of the GHs were found within the Shi *et al*. (2014; 67) dataset confirms that they are actively expressed within the rumen and not artefacts of the genome assembly.

Conclusions

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 highlights the extensive genomic variation found within *Butyrivibrio*, and to a lesser extent, *Pseudobutyrivibrio*. We also demonstrate the existence of outlier strains within the existing taxonomy in terms of phylogeny, GC content, genome size, and ANI and suggest the existence of a new *Butyrivibrio* species. The *Butyrivibrio* and *Pseudobutyrivibrio* genomes are also very open with very low % core genomes and high % accessory genomes. Despite genomic variation, taxa appear to retain broadly similar high-level functional profiles, and possess a number of GH isoforms that we hypothesise facilitate metabolic plasticity and resilience under dietary perturbations.

Methods

Genomes used in this study

Seventy genomes of *Butyrivibrio/ Pseudobutyrivibrio* isolates were obtained from the Hungate 1000 collection (JGI; 16) and one additional strain, *Pseudobutyrivibrio xylanivorans* MZ8 (obtained from the Rowett Research Institute, University of Aberdeen), was genome sequenced in this study (Table 1) and submitted to GenBank (BioProject number PRJNA563299). The sample was then sent for sequencing to MicrobesNG (https://microbesng.uk/), where it was sequenced on the Illumina HiSeq 2500 platform, using 2x250bp paired-end reads and with x30 coverage. Following sequencing, the data were put through MicrobesNG's standard analysis pipeline, which included strain identification by Kraken [68], *de novo* assembly of the reads by SPAdes [69]. All 71 genomes were re-annotated using Prokka V1.12 [70] via the Galaxy platform [71] with a similarity e-value cut-off of 1×10^{-6} to ensure analytical consistency. The 16S rDNA sequences for these genomes were obtained from the Prokka annotations.

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 [72], and a phylogenetic tree constructed using FastTree V2.1.10 [73], using default parameters. An additional tree was constructed using 40 gene markers as per Wu and Eisen (2008; 39) and Creevey *et al*. (2011; 40). Both trees were visualised by the Interactive Tree Of Life (iToL), changelog version 3.5.2 [74]. *Clostridium beijerinckii* NCIMB 8052 and *Lactobacillus acidophilus* NCFM were used as outgroups, and the tree was rooted for visualisation.

ANI was calculated using the PyANI script (available at https://github.com/widdowquinn/scripts/blob/master/bioinformatics/calculate_ani.py). Input sequences were in FASTA format, and were aligned using MUMmer (NUCmer). The comparisons were visualised by selecting for heatmap and dendogram output.

Pangenomics

Pangenomic analysis was done according to the species level groups found on the 40 marker tree, i.e. all strains listed as *B. fibrisolvens* underwent pangenomic analysis together, all *B. hungatei* strains underwent separate analysis, et cetera. Core and accessory genomic fragments were identified from the Prokka annotated genomic sequences (.ffn files) using Spine V0.3.1 (Ozer *et al*., 2014, http://vfsmspineagent.fsm.northwestern.edu/index_age.html; 29) with default parameters (clustering at 85% similarity and identifying core genes when present in 100% of the genomes analysed). A range of other defined parameters (70-100% similarity and present in 50-100% of genomes) were used to check which parameters are optimal to use beforehand. Accessory elements were visualised using ClustAGE V0.8 and ClustAGE Plot [29] for each group (as identified from the phylogenetic analysis described previously). The minimum accessory genomic element (AGE) size to represent in the ClustAGE Plot was set to 1500 bp due to file size restrictions when using ClustAGE Plot.

Core and accessory fasta files were combined into the core genome and the accessory genome for their respective taxa and uploaded to EggNOG [75] to determine genomic subsystem allocation. A stacked histogram was then made using these data to compare core and accessory functionality on a taxon level. Core and accessory gene GC% content were taken from the Spine statistics files for each run of Spine also.

Gene Evolution

Putative gene orthology within the 71 genomes was determined using OrthAgogue [31] 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} (this was also completed using an E-value cut-off of 10^{-5} and data obtained was similar; data not shown) and forced OrthoMCL [76] emulation. All other OrthAgogue parameters were default. The clusters identified by OrthAgogue were turned into binary data, indicating either the presence or absence of a cluster of orthologous and inparalogous genes in each strain, after which the lists of clusters were uploaded to UpSet [77] to visualise the intersections. All groups were selected to be visualised, and the number of intersections was set to 60.

Carbohydrate-active enzymes (CAZymes)

CAZymes were identified using the dbCAN metaserver [78] and annotated using the "mRNAs/CDSs/Metagenomes or short DNA seqs" option, running HMMER with default setting of E-Value < 1e-15, coverage > 0.35. GH sequences were extracted using Samtools V1.9 and each GH family was aligned using the Clustal Omega online server. A tree for each family was inferred using IQ tree V1.6.10, and visualised using iToL. Coloured ranges were put onto the trees using the iTOL template sheets to display clade colourings, and OrthAgogue and dbCAN annotations were compared to determine which GHs were orthologs, co-orthologs and inparalogs. These homolog-based annotations were denoted by the concentric coloured rings surrounding the trees, again using the iToL templates. Stacked histograms were produced showing only the most abundant GH families; in the case of the histogram with all 71 genomes displayed, and the histogram showing all 6 taxa, only GH families with over 100 total instances across all genomes were displayed. For the taxon specific histograms, this number was 10.

Metranscriptome analysis

To check expression of the identified CAZyme isoforms in the rumen, and check whether these genes are expressed *in vivo* conditions, twenty publicly available metatranscriptomic datasets were taken from National Centre for Biotechnology Information Sequence Read Archive, under the accession number SRA075938 [67]. Datasets were composed of 150 bp paired end reads from the illumina Hiseq 2000 sequencer [67]. Fastq files were processed with multiqc [79] and reads were trimmed from 150 bp to 110 bp using trimmomatic software version 0.36 [80]. Reads were aligned to the Hungate rumen genome dataset using bowtie2 version 2.3.0 [81] using the settings "–very-sensitive-local" which allowed soft trimming of the reads and a relaxed alignment; and "-k 497". This produced SAM files, which were converted to BAM files using SAMtools version 1.9 [82]. SAMtools version 1.9 was used to filter all and the best alignment position for each read using the flag option "-F 260". For each of the resultant 20 BAM files FeatureCounts (from the subread package version 2.0.0) [83] was used to calculate the number of reads that align within the boundaries of every predicted gene in the Hungate genomes. Read counts were then converted into RPKM values. Finally, the RPKM values of the CAZyme gene haplotypes were extracted from the entire expression count table. If a gene was found in a metatranscriptome (expressed to any degree) then it was visualised on the outside of the iToL GH trees as a black square.

Declarations

Ethics approval and consent to participate

Not Applicable

Availability of data and materials

Pseudobutyrivibrio xylanivorans MZ8 was genome sequenced in this study and the genome submitted to GenBank (BioProject number PRJNA563299).

Competing interests

All authors declare no financial or non-financial competing interests.

Funding

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

Authors' contributions

SAH and SP conceptualized the research and led the project. TS, KST, FGS and CJC helped SP with the computational analyses and discussions regarding project direction. SJC performed some computation analyses. SP, SAH, CJC and TS drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Authors information

Not Applicable

Tables

Table 1 - Strains used in this study, separated into their respective species-level taxa based on a 40 marker gene phylogenetic analysis and pre-existing taxonomy.

Phylogenetic clade	Strain ID	Genome size (Mbp)	Number of genes
B. fibrisolvens	AB2020	4.74	3920
	AR40	5.01	4148
	D1	4.84	4030
	FE2007	4.32	3603
	MC2013	3.81	3146
	MD2001	4.74	3889
	NC3005	3.87	3273
	ND3005	4.46	3676
	TB	4.41	3717
	WTE3004	4.66	3833
	YRB2005	4.65	3856
B. hungatei	AE2005	3.71	3375
	AE3003	3.46	3272
	INlla18	3.01	2959
	INlla21	3.34	3146
	JK615	3.39	3037
	LB2008	3.72	3408
	M55	3.67	3336
	MB2003	3.24	2873
	NK4A153	3.37	3053
	XBD2006	3.40	3070
	YAB3001	4.58	4079
B. proteoclasticus	AE2015	3.71	3169
	AE2032	3.69	3247
	AE3009	4.22	3480
	B316	3.86	3231
	FCS006	3.79	3253
	FCS014	4.09	3573
	FD2007	3.88	3243
	INlla14	4.19	3684
	MC2021	4.36	3569
	NC2007	4.00	3528
	P6B7	4.17	3708
	P18	3.93	3461
	Su6	3.69	3119
	VCB2001	4.02	3529
	VCB2006	3.95	3378

	XBB1001	3.91	3324
	XPD2006	3.89	3439
Butyrivibrio sp.	AC2005	5.12	4625
	AD3002	4.32	3824
	AE3004	4.47	3947
	AE3006	4.15	3545
	FC2001	4.57	4045
	LC3010	4.59	3964
	MB2005	4.18	3606
	NC2002	3.41	2959
	OB235	4.73	4249
	VCD2006	4.70	3976
	WCD2001	4.28	3795
	WCD3002	4.25	3666
	WCE2006	4.53	3909
	XPD2002	4.44	3926
P. ruminis	A12-1	3.02	2718
	ACV-9	2.76	2512
	AD2017	2.91	2561
	C4	2.92	2688
	CF1b	3.27	2918
	HUN009	2.80	2526
	JW11	3.06	2757
	LB2011	2.81	2552
	MD2005	3.03	2740
	OR37	3.56	3098
P. xylanivorans	Sp 49	3.40	3046
	ACV-2	3.66	3253
	AR14	3.05	2769
	Bu21	3.21	2883
	Mz5	3.42	3052
	Mz8	2.96	2663
	NOR37	2.94	2661
	YE44	3.24	2896

Table 2 - Phylogenetic groups used in this study. The total number of core and accessory genes per group is listed, as well as the total pangenome and and the core/pangenome ratio for core definitions of both 90% and 100%. Pangenome analysis was performed by Spine (Ozer *et al.*, 2014). http://eggnogdb.embl.de/#/app/emapper).

Phylogenetic clade	Total core genes (90%)	Total accessory genes (90%)	Total genes in pangenome (90%)	Core/pangenome ratio (90%)	Total core genes (100%)	Total accessory genes (100%)	Total genes in pangenome (100%)	Core/pangenome ratio x 100 (100%)
B. fibrisolvens	1036	40999	42035	2.46	380	41490	41870	0.91
B. hungatei	1493	35220	36713	4.07	1118	35488	36606	3.05
B. proteoclasticus	2057	57764	59821	3.44	1704	57963	59667	2.86
Butyrivibrio sp.	2558	53342	55900	4.58	1868	53663	55531	3.36
P. ruminis	2960	25701	28661	10.33	2246	26438	28684	7.83
P. xylanivorans	1874	22571	24445	7.67	1874	22571	24445	7.67

References

1. Bapteste E, Boucher Y. 2009. Epistemological impacts of horizontal gene transfer on classification in microbiology. *Methods Mol Biol* 532:55-72. doi: 10.1007/978-1-60327-853-9_4.
2. Papke R. 2009. A critique of prokaryotic species concepts. *Methods Mol Biol* 532:379-95. doi: 10.1007/978-1-60327-853-9_22.
3. Wayne L, Moore W, Stackebrandt E, et al. 1987. Report of the Ad Hoc Committee on reconciliation of approaches to bacterial systematics. *Int J Sys Evol Microbiol* 37:463-464. 0020-7713/87/040463-02$02.00/0.
4. Schleifer K. 2009. Classification of Bacteria and Archaea: Past, present and future. *Sys Appl Microbiol* 32:533-542. doi: 10.1016/j.syapm.2009.09.002.
5. Woese C, Stackebrandt E, Macke T, et al. 1985. A phylogenetic definition of the major eubacterial taxa. *Sys Appl Microbiol* 6:143-151. doi: 10.1016/s0723-2020(85)80047-3.
6. Zhang W, Du P, Zheng H, et al. 2014. Whole-genome sequence comparison as a method for improving bacterial species definition. *J Gen Appl Microbiol* 60:75-78. doi: 10.2323/jgam.60.75.
7. Mende D, Sunagawa S, Zeller G, et al. 2013. Accurate and universal delineation of prokaryotic species. *Nature Methods* 10:881-884. doi: 10.1038/nmeth.2575.
8. Rossi-Tamisier M, Fournier P, Benamar S, et al. 2015. Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. *Int J Sys Evol Microbiol* 65:1929-1934. doi: 10.1099/ijs.0.000161.
9. Konstantinidis K, Tiedje J. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Nat Acad Sci* 102:2567-2572. doi: 10.1073/pnas.0409727102.
10. Cowan S. 1971. Sense and Nonsense in Bacterial Taxonomy. *J Gen Microbiol* 67:1-8.
11. Rouli L, Merhej V, Fournier P, Raoult, D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* 7:.72-85. doi: 10.1016/j.nmni.2015.06.005.
12. Moldovan M, Gelfand M. 2018. Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus Front Microbiol* 9. doi: 10.3389/fmicb.2018.00428.
13. Goryunov D, Nagaev B, Nikolaev M, et al. 2015. Moss phylogeny reconstruction using nucleotide pangenome of complete mitogenome sequences. *Biochem* 80:1522-1527. doi: 10.1134/S0006297915110152.
14. García-López M, Meier-Kolthoff J, Tindall B, et al. 2019. Analysis of 1,000 Type-Strain Genomes Improves taxonomic classification of Bacteroidetes. *Front Microbiol.* doi.org/10.3389/fmicb.2019.02083
15. Ricard G, McEwan N, Dutilh B,. et al. 2006. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* 7:22. https://doi.org/10.1186/1471-2164-7-22.
16. Seshadri R, Leahy S, Attwood G, et al. 2018. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nature Biotechnol* 36:.359-367. https://doi.org/10.1038/nbt.4110.
17. Henderson G, Cox F, Ganesh S, et al. 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci Reports* 5(1). org/10.1038/srep14567.
18. Bryant, M, Small, N.,1956. The anaerobic monotrichous butyric acid-producing curved rod-shaped bacteria of the rumen. *J Bacteriol* 72:16-21.

19. Bryant M. 1986. Bergey's Manual of Systematic Bacteriology. Baltimore: Williams and Wilkins.
20. Cheng K, Costerton J. 1977. Ultrastructure of *Butyrivibrio fibrisolvens*: a Gram-positive bacterium? *J Bacteriol* 129:1506-1512.
21. Hespell R, Wolf R, Bothast R. 1987. Fermentation of xylans by *Butyrivibrio fibrisolvens* and other ruminal bacteria. *Appl Environ Microbiol* 53:2849-2853.
22. Stewart R, Auffret M, Warr A, et al. 2019. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnol* 37:953-961. https://doi.org/10.1038/s41587-019-0202-3.
23. Kopečný J, Zorec M, Mrázek J, et al. 2003. *Butyrivibrio hungatei* nov. and *Pseudobutyrivibrio xylanivorans* sp. nov., butyrate-producing bacteria from the rumen. *Int J Systematic Evol Microbiol* 53:201-209. DOI: 10.1099/ijs.0.02345-0.
24. Moore W, Johnson J, Holdeman L. 1976. Emendation of Bacteroidaceae and *Butyrivibrio* and descriptions of *Desulfomonas* nov. and ten new species in the genera *Desulfomonas*, *Butyrivibrio*, *Eubacterium, Clostridium*, and *Ruminococcus*. *Int J Sys Bacteriol* 26:238-252.
25. Moon C, Pacheco D, Kelly W, et al. 2008. Reclassification of *Clostridium proteoclasticum* as *Butyrivibrio proteoclasticus* nov., a butyrate-producing ruminal bacterium. *Int J Sys Evol Microbiol* 58:2041-2045. DOI: 10.1099/ijs.0.65845-0.
26. Van Gylswyk N, Hippe H, Rainey F. 1996. *Pseudobutyrivibrio ruminis* nov., sp. nov., a butyrate-producing bacterium from the rumen that closely *resembles Butyrivibrio fibrisolvens* in phenotype. *Int J Sys Bacteriol* 46:559-563.
27. Palevich N, Kelly W, Leahy S, et al. 2019. Comparative genomics of rumen *Butyrivibrio* uncovers a continuum of polysaccharide-degrading capabilities. *Appl Environ Microbiol* doi:1128/AEM.01993-19.
28. Pritchard L, Glover R, Humphris S, et al. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 8:12-24.
29. Ozer E, Allen J, Hauser A. 2014. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* 15:737. https://doi.org/10.1186/1471-2164-15-737.
30. Maistrenko O, Mende D, Luetge M, et al. 2020. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J* 14:1247-1259. https://doi.org/10.1038/s41396-020-0600-z.
31. Ekseth O, Kuiper M, Mironov V. 2013. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* 30:734-736. DOI: 1093/bioinformatics/btt582.
32. Kelly W, Leahy S, Altermann E, et al. 2010. The glycobiome of the rumen bacterium *Butyrivibrio proteoclasticus* B316T highlights adaptation to a polysaccharide-rich environment. *PLoS ONE* 5:e11942. https://doi.org/10.1371/journal.pone.0011942.
33. Cotta M, Hespell R. 1986. Proteolytic activity of the ruminal bacterium *Butyrivibrio fibrisolvens*. *Appl Environ Microbiol* 52:51-58.
34. Paillard D, McKain N, Chaudhary L, et al. 2007. Relation between phylogenetic position, lipid metabolism and butyrate production by different *Butyrivibrio*-like bacteria from the rumen. *Antonie van Leeuwenhoek* 9:417-422. DOI: 1007/s10482-006-9121-7.
35. Palevich N, Kelly W, Leahy S, et al. 2017. The complete genome sequence of the rumen bacterium *Butyrivibrio hungatei Standards Genomic Sci* 12(1). doi: 10.1186/s40793-017-0285-8.
36. Hutchison C, Chuang R, Noskov V, et al. 2016. Design and synthesis of a minimal bacterial genome. Science 351:6253-6253. DOI: 1126/science.aad6253.
37. Bond J, Dunne J, Kwan F, et al. 2012. Carbohydrate transporting membrane proteins of the rumen bacterium, *Butyrivibrio proteoclasticus*. *J Proteomics* 75:3138-3144. doi: 10.1016/j.jprot.2011.12.013.
38. Marounek M, Petr O. 1995. Fermentation of glucose and xylose in ruminal strains of *Butyrivibrio fibrisolvens*. *Lett Appl Microbiol* 21:272-276. doi: 10.1111/j.1472-765x.1995.tb01058.x.
39. Wu M, Eisen J. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9:R151. https://doi.org/10.1186/gb-2008-9-10-r151.
40. Creevey C, Doerks T, Fitzpatrick D, et al. 2011. Universally Distributed Single-copy genes indicate a constant rate of horizontal transfer. *PLoS ONE* 6:e22099. https://doi.org/10.1371/journal.pone.0022099.
41. Kasperowicz A, Stan-Glasek K, Guczynska W, et al. 2009. Sucrose phosphorylase of the rumen bacterium *Pseudobutyrivibrio ruminis* strain A12-1. *J Appl Microbiol* 107:.812-820. DOI: 1111/j.1365-2672.2009.04257.x.
42. Lane D, Pace B, Olsen G, et al. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Nat Acad Sci* 82:6955-6959. DOI: 1073/pnas.82.20.6955.
43. Acinas S, Marcelino L, Klepac-Ceraj V, et al. 2004. Divergence and redundancy of 16S rRNA Sequences in genomes with multiple rrn operons. *J Bacteriol* 186:2629-2635. DOI: 1128/jb.186.9.2629-2635.2004.

44. Beumer A, Robinson J. 2005. A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Appl Environ Microbiol* 71:8301-8304. doi: 10.1128/AEM.71.12.8301-8304.2005.
45. McNally A, Oren Y, Kelly D, et al. 2016. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genetics* 12:e1006280. https://doi.org/10.1371/journal.pgen.1006280.
46. McInerney J, McNally A, O'Connell M. 2017. Why prokaryotes have pangenomes. *Nature Microbiol* 2(4). org/10.1038/nmicrobiol.2017.40.
47. Gabaldón T, Koonin E. 2013. Functional and evolutionary implications of gene orthology. *Nature Rev Genetics* 14:360-366. https://doi.org/10.1038/nrg3456.
48. Hudman J, Gregg K. 1989. Genetic diversity among strains of bacteria from the rumen. *Current Microbiol* 19:313-318. https://doi.org/10.1007/BF01570107.
49. Mannarelli B. 1988. Deoxyribonucleic acid relatedness among strains of the species *Butyrivibrio fibrisolvens*. *Int J Sys Bacteriol* 38:340-347. 0020-7713f88/040340-08$02.OO/O.
50. Lassalle F, Périan S, Bataillon T, et al. 2015. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. *PLoS Genetics* 11:e1004941. https://doi.org/10.1371/journal.pgen.1004941.
51. Pozzoli U, Menozzi G, Fumagalli M, et al. 2008. Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol Biol* 8: p.99. https://doi.org/10.1186/1471-2148-8-99.
52. Huang S, Zhang S, Jiao N, et al. 2015. Comparative genomic and phylogenomic analyses reveal a conserved core genome shared by estuarine and oceanic Cyanopodoviruses. *PLoS ONE* 10:e0142962. https://doi.org/10.1371/journal.pone.0142962.
53. Hiratani I, Leskovar A, Gilbert D. 2004. Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proc Nat Acad Sci* 101;16861-16866. https://doi.org/10.1073/pnas.0406687101.
54. Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of Selection upon genomic GC-content in bacteria. *PLoS Genetics* 6:1-9. https://doi.org/10.1371/journal.pgen.1001107.
55. Wuitschick J, Karrer K. 1999. Analysis of genomic G + C Content, codon usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*. *J Euk Microbiol* 46:239-247. DOI: 1111/j.1550-7408.1999.tb05120.x.
56. Mansur M, Cabello C, Hernàndez L, et al. 2005. Multiple gene copy number enhances insulin precursor secretion in the yeast *Pichia pastoris*. *Biotechnol Lett* 27:339-345. https://doi.org/10.1007/s10529-005-1007-7.
57. Couturier E, Rocha E. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol* 59:1506-1518. DOI: 1111/j.1365-2958.2006.05046.x
58. Walsh J, Stephan W. 2008. Multigene Families: Evolution. Encyclopedia of Life Sciences, Nature Publishing Group.
59. Rubino F, Carberry CM, Waters S., et al. 2017. Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. *ISME J* 11:932-944. https://doi.org/10.1038/ismej.2016.172.
60. Dagan T, Martin W. 2006. The tree of one percent. *Genome Biol* 7:118. https://doi.org/10.1186/gb-2006-7-10-118.
61. Lombard V, Golaconda Ramulu H, et al, 2013. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42:D490-D495. DOI: 1093/nar/gkt1178.
62. Intra J, Pavesi G, Horner D. 2008. Phylogenetic analyses suggest multiple changes of substrate specificity within the glycosyl hydrolase 20 family. *BMC Evol Biol* 8:214. https://doi.org/10.1186/1471-2148-8-214.
63. Ferrer M, Ghazi A, Beloqui A, et al. 2012. Functional metagenomics unveils a multifunctional glycosyl hydrolase from the family 43 catalysing the breakdown of plant polymers in the calf rumen. *PLoS ONE* 7:e38134. https://doi.org/10.1371/journal.pone.0038134.
64. Nei M, Rooney A. 2005. Concerted and birth-and-death evolution of multigene families. *Annual Rev Genetics* 39:121-152. doi: 10.1146/annurev.genet.39.073003.112240.
65. Mertz B, Gu X, Reilly P. 2009. Analysis of functional divergence within two structurally related glycoside hydrolase families. *Biopolymers* 91:478-495. doi: 10.1002/bip.21154.
66. Ohta T. 2008. Gene Families: Multigene families and superfamilies. Encyclopedia of Life Sciences, John Wiley & Sons Ltd.
67. Shi W, Moon C, Leahy S, et al. 2014. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res* 24:1517-1525. doi: 10.1101/gr.168245.113.
68. Bankevich A, Nurk S, Antipov D, , et al. 2012. SPAdes: A New genome assembly algorithm and Its applications to single-cell sequencing. *J Comp Biol* 19:455-477. doi: 10.1089/cmb.2012.0021.

69. Wood D, Salzberg S. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15:.R46. https://org/10.1186/gb-2014-15-3-r46.
70. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-2069. DOI: 1093/bioinformatics/btu153.
71. Afgan E, Baker D, van den Beek M, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 44:W3-W10. doi: 10.1093/nar/gkw343.
72. Cole J, Wang Q, Fish J, et al. 2013. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:D633-D642. doi: 10.1093/nar/gkt1244.
73. Price M, Dehal P, Arkin A. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
74. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44;W242-W245. doi: 10.1093/nar/gkw290.
75. Huerta-Cepas J, Forslund K, Coelho L, et al. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 34:2115-2122. DOI: 1093/molbev/msx148.
76. Li L, Stoeckert C, Roos D. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178-2189. DOI: 1101/gr.1224503.
77. Lex A, Gehlenborg N, Strobelt H, et al. 2014. UpSet: Visualization of intersecting sets. *IEEE Trans Visualization Comp Graphics* 20:1983-1992. DOI: 1109/TVCG.2014.2346248.
78. Yin Y, Mao X, Yang J, et al. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 40:W445-W451. DOI: 1093/nar/gks479.
79. Ewels P, Magnusson M, Lundin, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single *Bioinformatics* 32:3047-3048. doi: 10.1093/bioinformatics/btw354.
80. Bolger A, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30:.2114- doi: 10.1093/bioinformatics/btu170.
81. Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357-359. https://doi.org/10.1038/nmeth.1923.
82. Li H, Handsaker B, Wysoker A, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078-2079. DOI: 1093/bioinformatics/btp352.
83. Liao Y, Smyth G, Shi W. 2013. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923-930. DOI: 1093/bioinformatics/btt656.

Figures

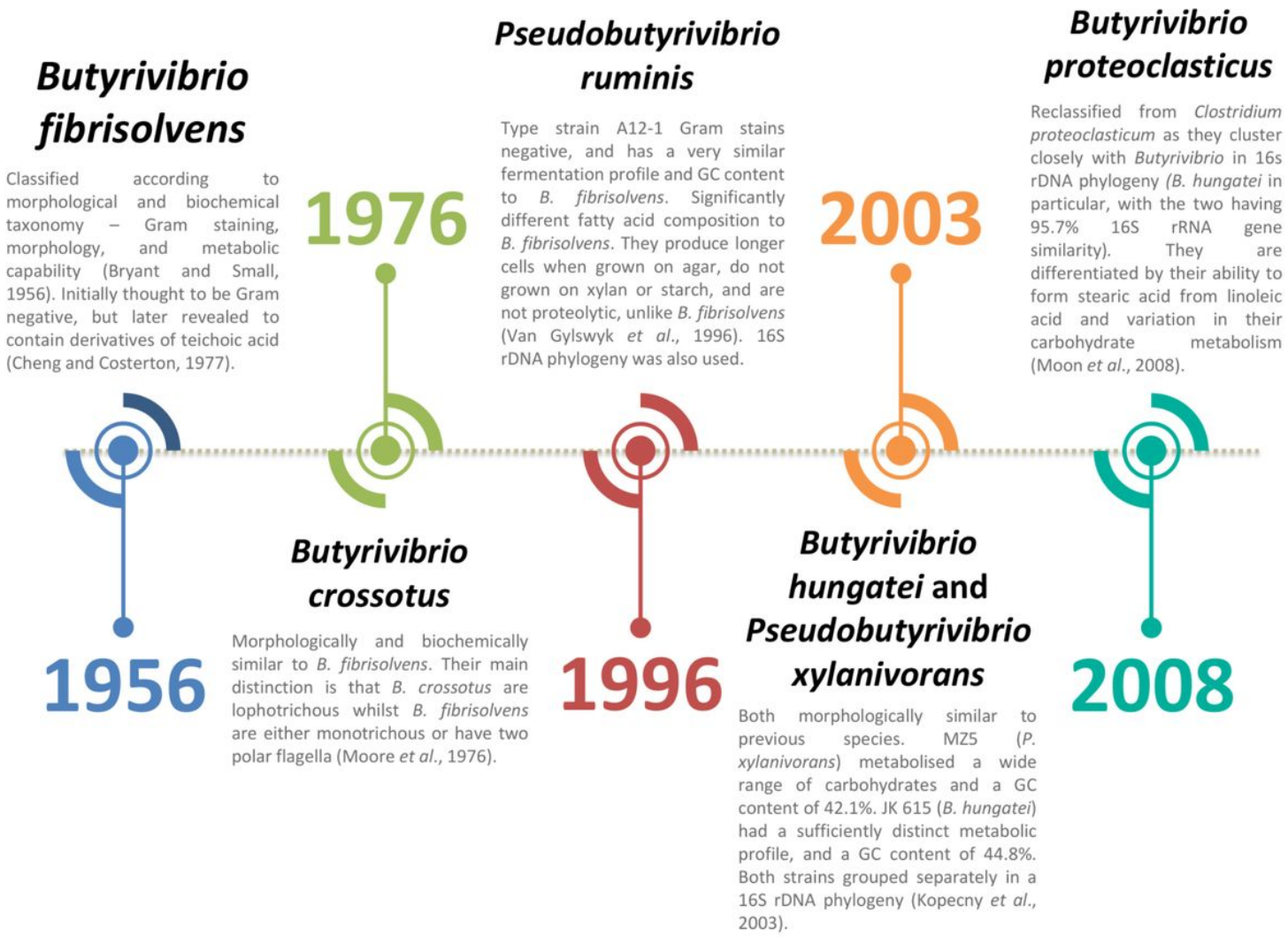

Figure 1

Chronological identification and classification of Butyrivibrio and Pseudobutyrivibrio.

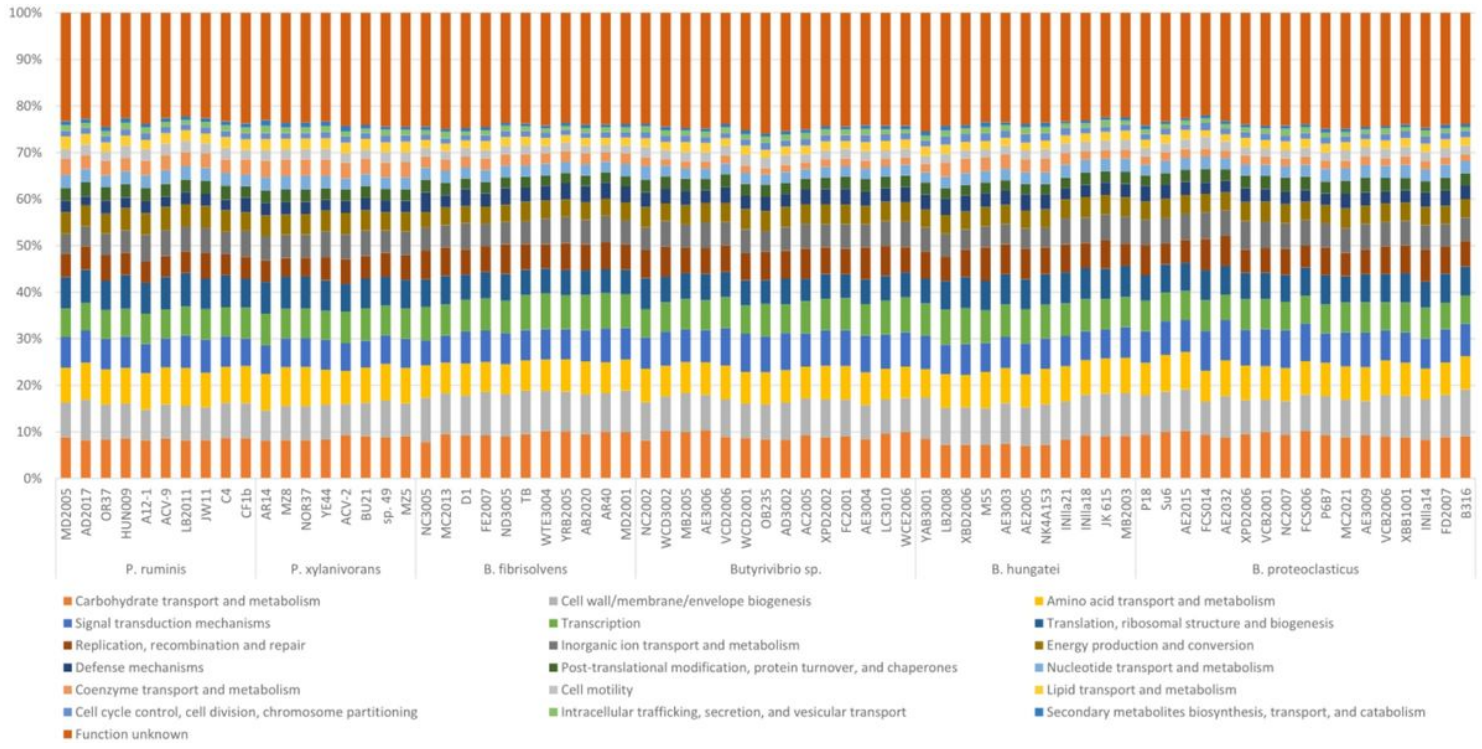

Figure 2

Functional annotation of the 71 Butyrivibrio and Pseudobutyrivibrio genomes used in this study. Gene functionality is sorted by colour, as indicated by the key. Annotation was performed using EggNOG (75; http://eggnogdb.embl.de/#/app/emapper).

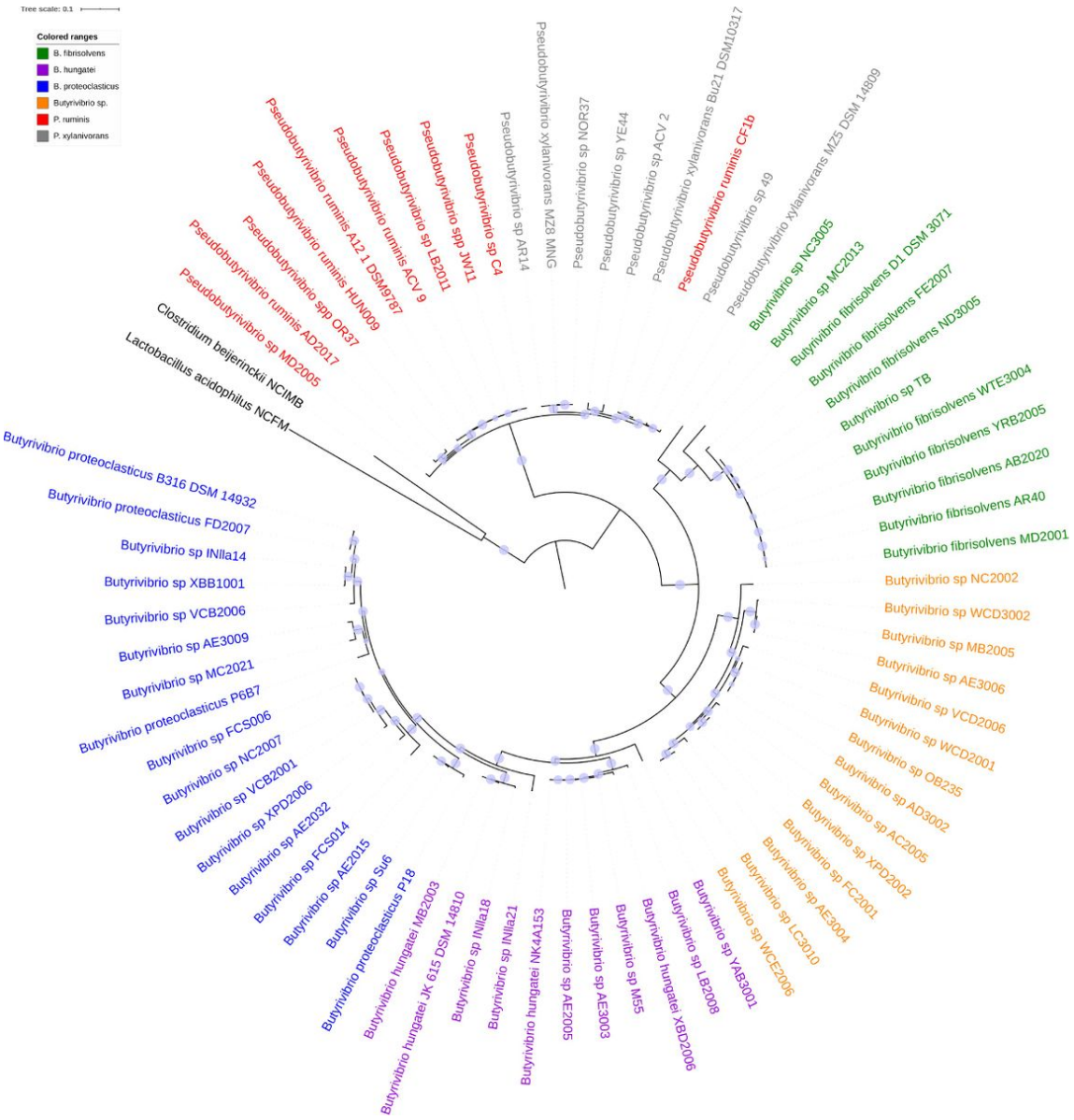

Figure 3

Phylogenetic tree showing the relatedness of 71 strains within the genera Butyrivibrio and Pseudobutyrivibrio based on 40 marker gene sequence comparisons. The scale bar represents 0.1 substitutions per nucleotide position, and differences between sequences are indicated by the branch lengths. Clostridium beijerinckii NCIMB 8052 and Lactobacillus acidophilus NCFM were used as outgroups to root the tree by. Colours denote the current species taxonomic assignments, as per the key. Bootstrap values are indicated by the blue circles.

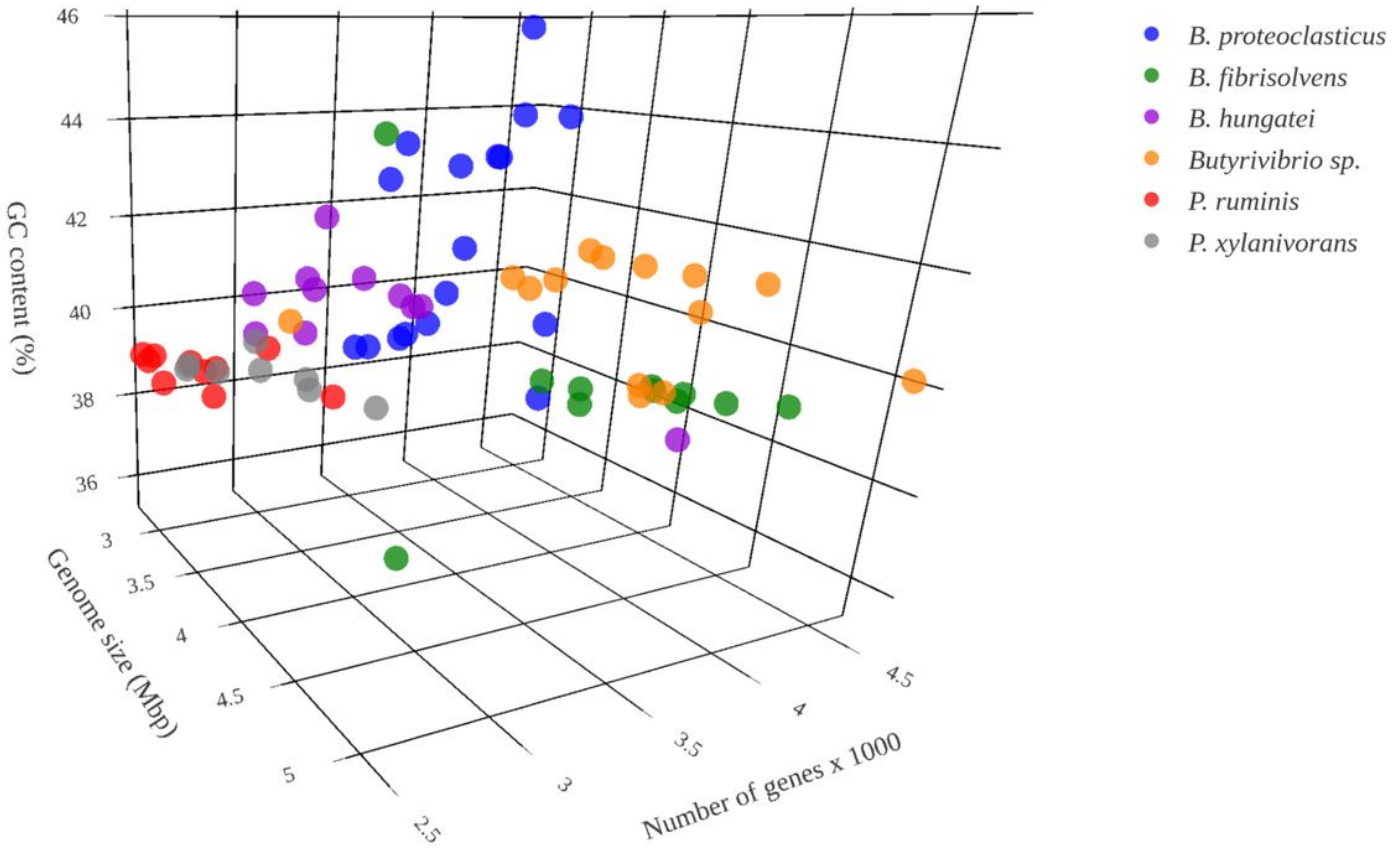

Figure 4

3D scatter plot of GC content (%), genome size (Mbp), and number of genes (x 1000) for 71 strains of Butyrivibrio and Pseudobutyrivibrio. Colours indicate the groups determined by classical taxonomy and 40 marker phylogeny, as indicated by the key.

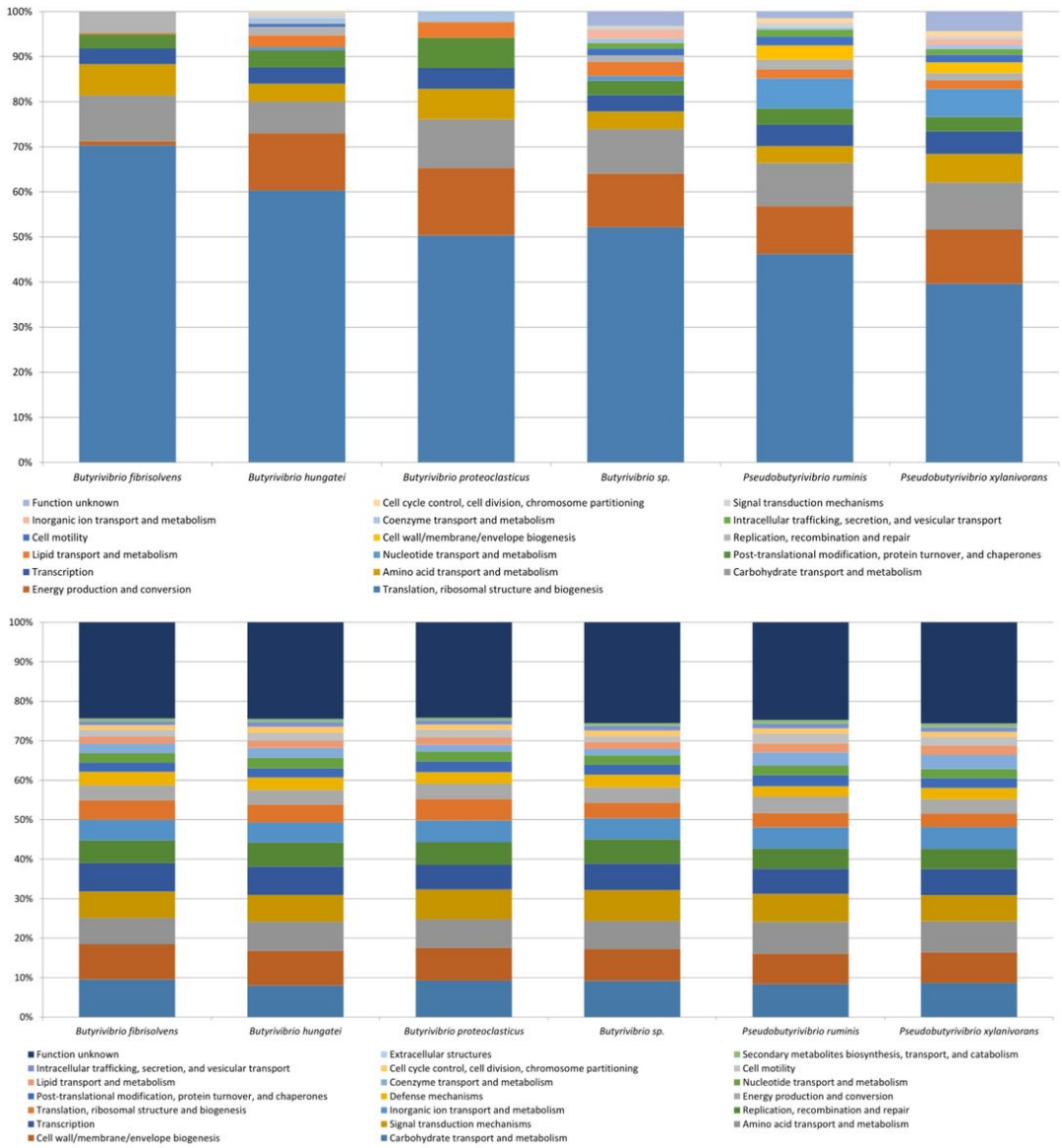

Figure 5

Functional annotations of Butyrivibrio and Pseudobutyrivibrio groups A. core and B. accessory genomes. Gene functionality is sorted by colour, as indicated by the key. Annotation was performed using EggNOG (75; http://eggnogdb.embl.de/#/app/emapper).

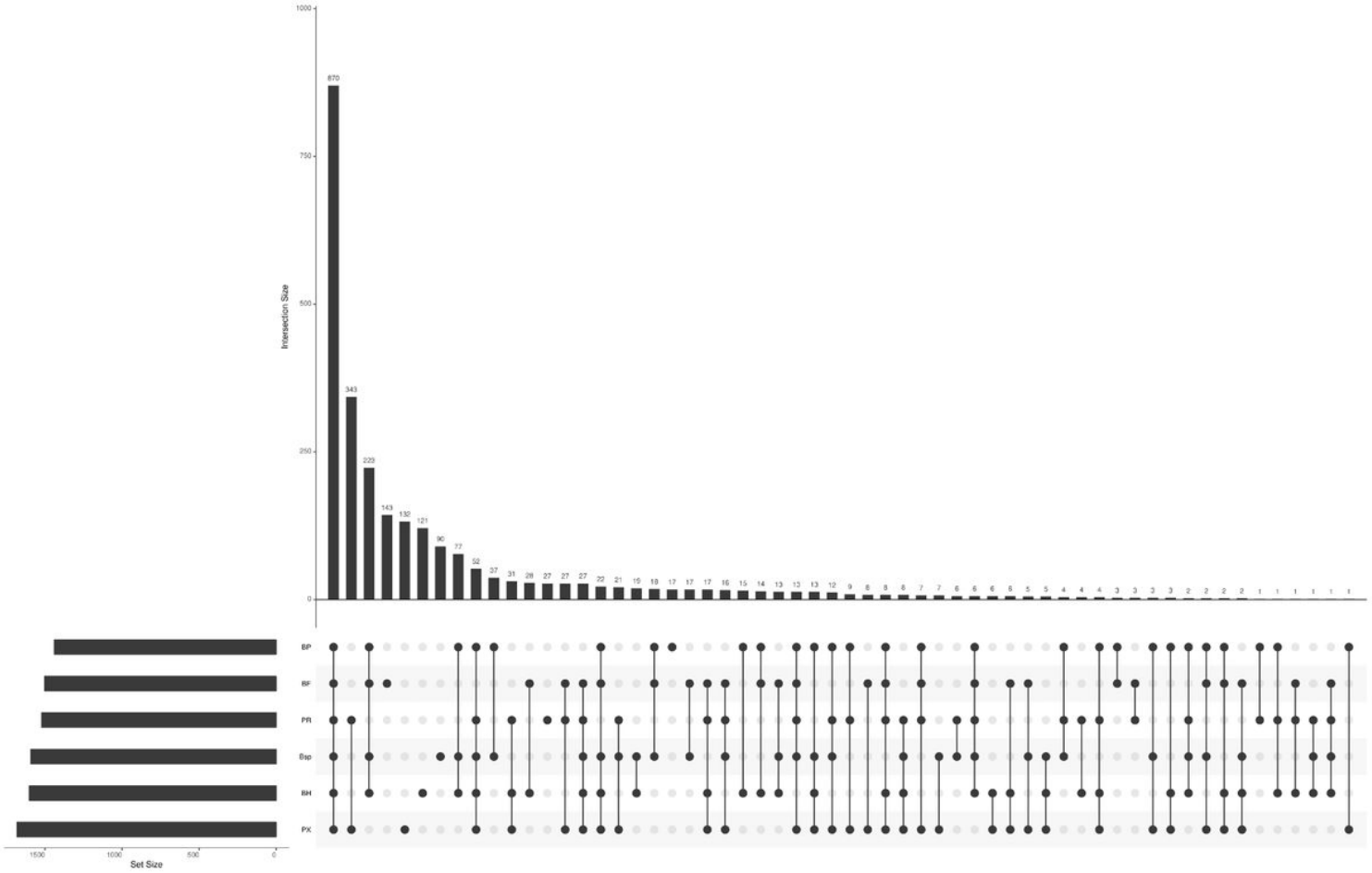

Figure 6

UpSet plot showing orthologous gene cluster intersections across six clade/species level groups (Butyrivibrio proteoclasticus, Butyrivibrio fibrisolvens, Pseudobutyrivibrio ruminis, Butyrivibrio sp., Butyrivibrio hungatei, and Pseudobutyrivibrio xylanivorans). Intersections are limited to 60, and are denoted by the corresponding bar.

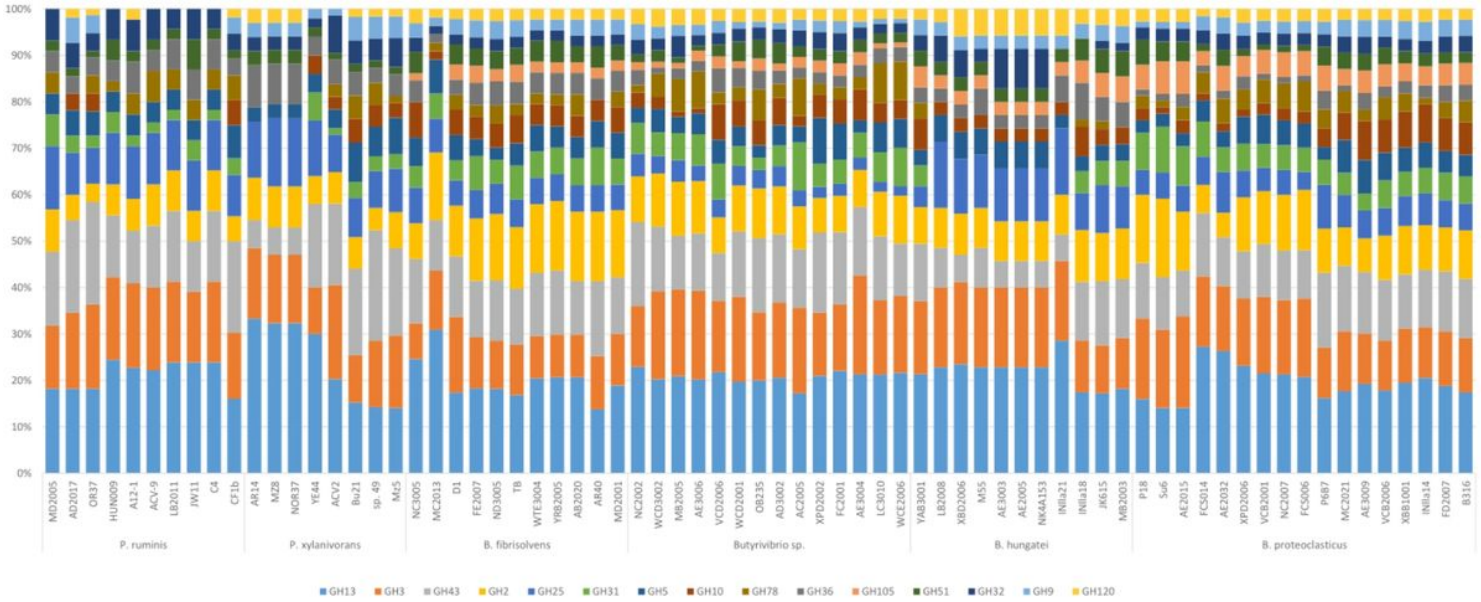

Figure 7

Proportions of the most abundant glycosyl hydrolase (GH) families found in 71 strains of Butyrivibrio and Pseudobutyrivibrio. GH families annotated by dbCan metaserver [78].

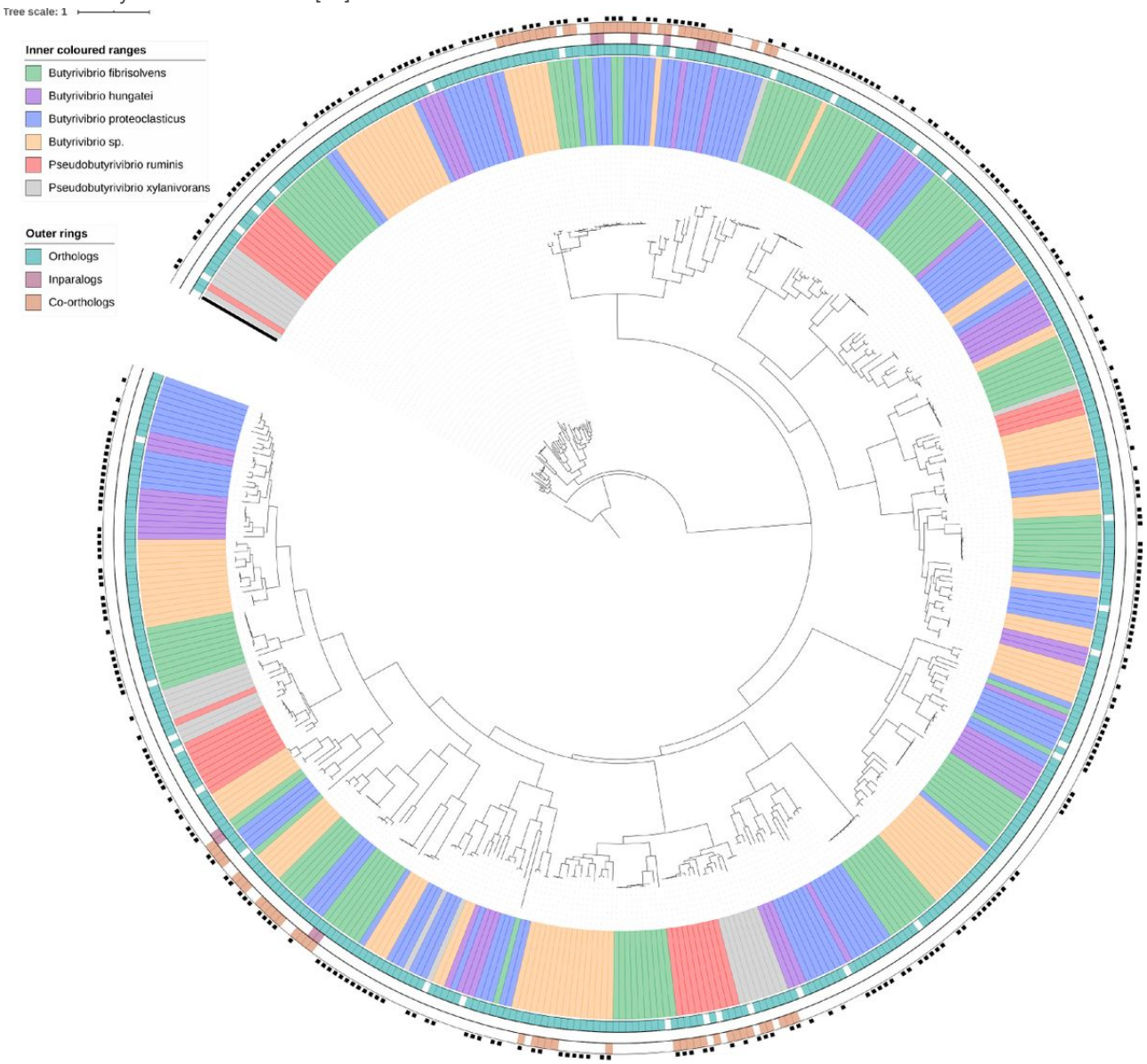

Figure 8

Phylogenetic tree showing the relatedness of all Glycosyl Hydrolase family 2 genes found in all 71 strains used in this analysis. Colours denote the clade groupings determined by 40 marker phylogeny, and the outer concentric rings denote genes annotated by OrthoAgogue as orthologs, inparalogs and co-orthologs, as denoted by the keys. The presence of a black square on the outermost layer indicates that that gene was found to be present in the Shi et al. (2014; 29) metatranscriptome dataset. The tree is rooted using a β-galactosidase large subunit sequence from Lactobacillus acidophilus NCFM, which is coloured in black.

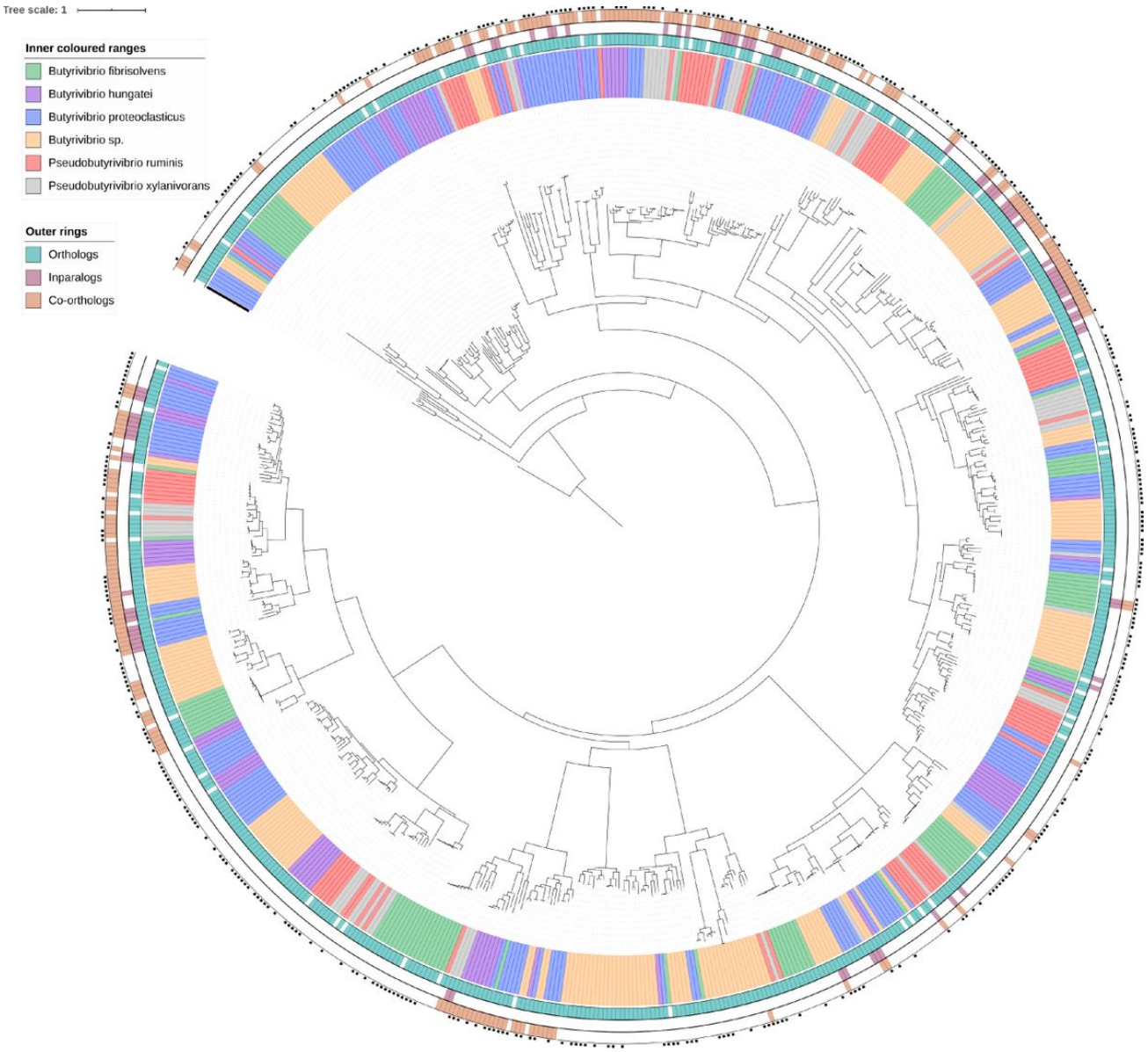

Figure 9

Phylogenetic tree showing the relatedness of all Glycosyl Hydrolase family 3 genes found in all 71 strains used in this analysis. Colours denote the clade groupings determined by 40 marker phylogeny, and the outer concentric rings denote genes annotated by OrthoAgogue as orthologs, inparalogs and co-orthologs, as denoted by the keys. The presence of a black square on the outermost layer indicates that that gene was found to be present in the Shi et al. (2014; 29) metatranscriptome dataset. The tree is rooted using a β-N-acetylhexosaminidase sequence from Lactobacillus acidophilus NCTC13720, which is coloured in black.

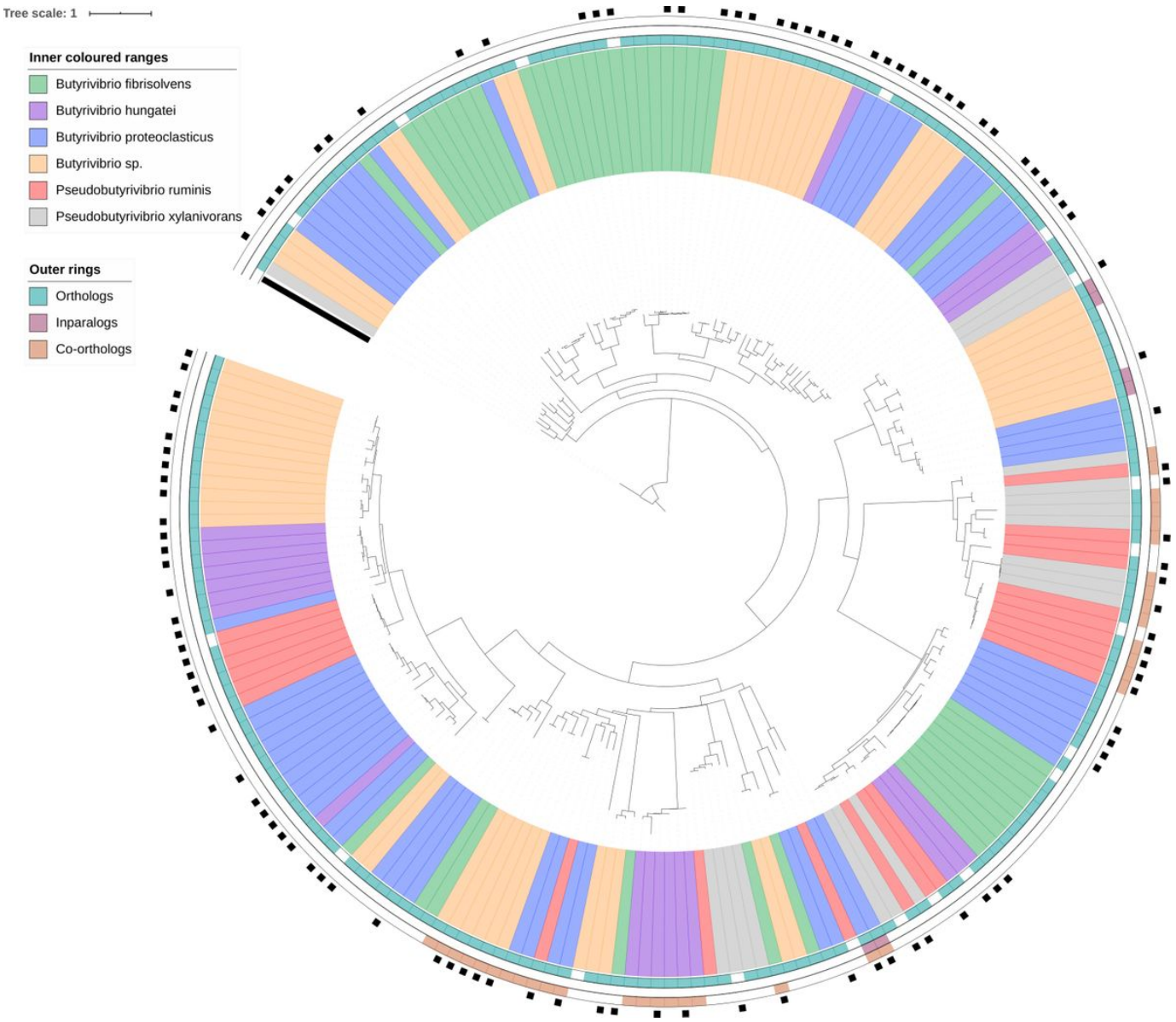

Figure 10

Phylogenetic tree showing the relatedness of all Glycosyl Hydrolase family 5 genes found in all 71 strains used in this analysis. Colours denote the clade groupings determined by 40 marker phylogeny, and the outer concentric rings denote genes annotated by OrthoAgogue as orthologs, inparalogs and co-orthologs, as denoted by the keys. The presence of a black square on the outermost layer indicates that that gene was found to be present in the Shi et al. (2014; 29) metatranscriptome dataset. The tree is rooted using a β-glucosidase sequence from Lactobacillus mucosae LM1, which is coloured in black.

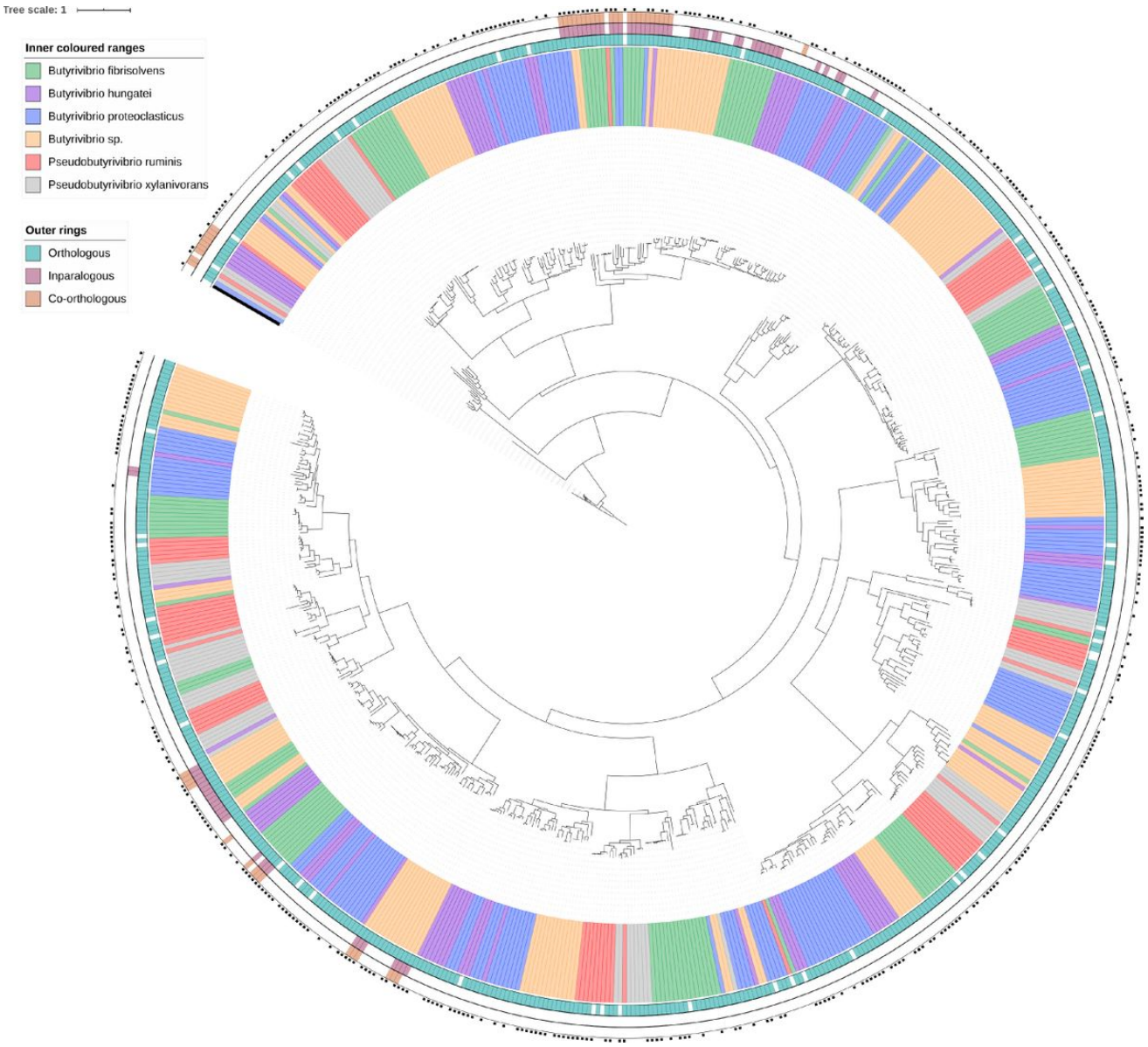

Figure 11

Phylogenetic tree showing the relatedness of all Glycosyl Hydrolase family 13 genes found in all 71 strains used in this analysis. Colours denote the clade groupings determined by 40 marker phylogeny, and the outer concentric rings denote genes annotated by OrthoAgogue as orthologs, inparalogs and co-orthologs, as denoted by the keys. The presence of a black square on the outermost layer indicates that that gene was found to be present in the Shi et al. (2014) metatranscriptome dataset. The tree is rooted using a sucrose phosphorylase sequence from Lactobacillus acidophilus NCFM, which is coloured in black.

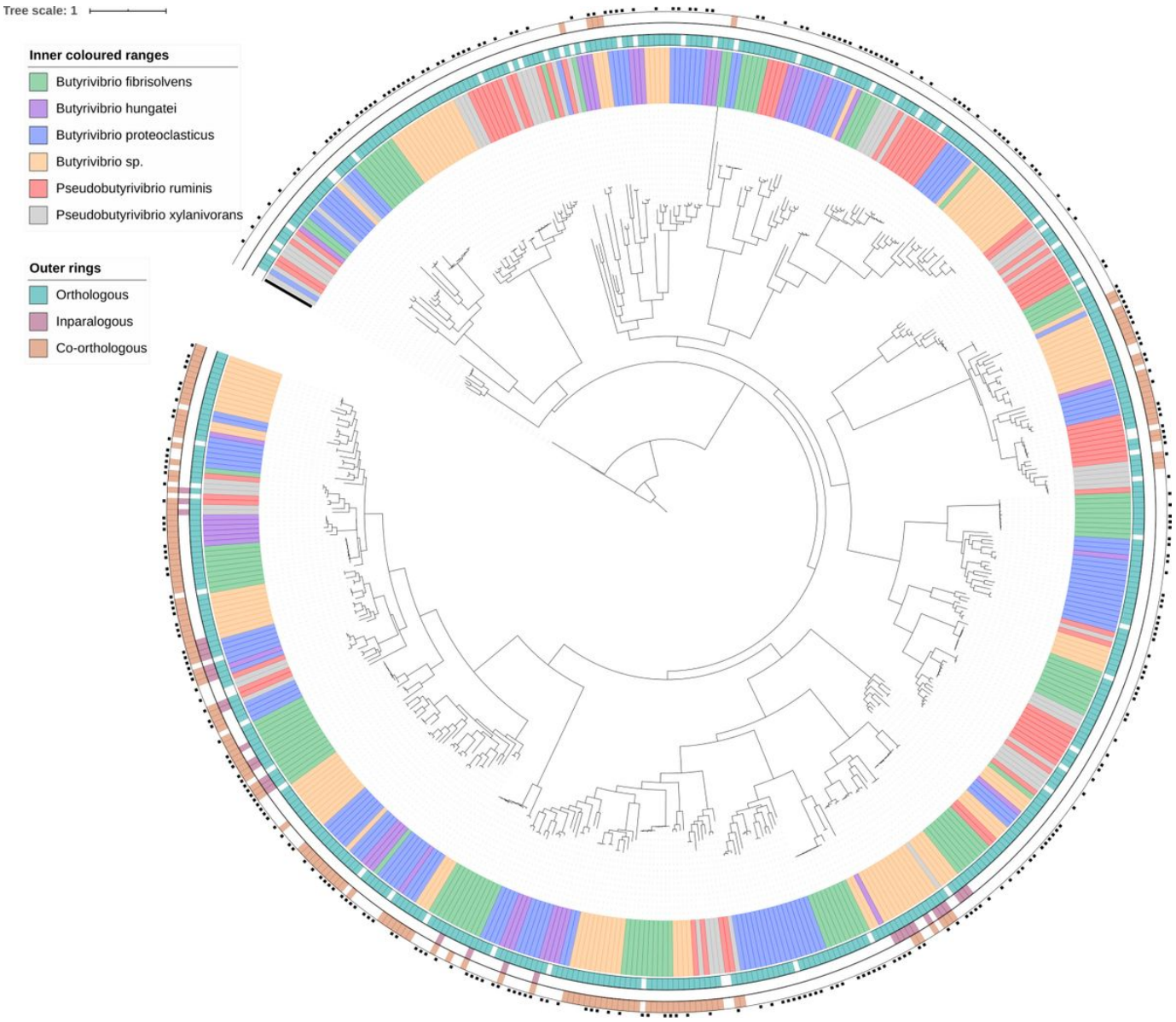

Figure 12

Phylogenetic tree showing the relatedness of all Glycosyl Hydrolase family 43 genes found in all 71 strains used in this analysis. Colours denote the clade groupings determined by 40 marker phylogeny, and the outer concentric rings denote genes annotated by OrthoAgogue as orthologs, inparalogs and co-orthologs, as denoted by the keys. The presence of a black square on the outermost layer indicates that that gene was found to be present in the Shi et al. (2014) metatranscriptome dataset. The tree is rooted using a β-xylosidase sequence from Lactobacillus mucosae LM1, which is coloured in black

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Excel1S01GHRPKM.xlsx
- Supplementarydata.pdf