Phylogenetic systematics of Butyrivibrio and Pseudobutyrivibrio pure culture and metagenomically assembled genomes suggest existence of 59 genera and 75 species, alongside possession of open pangenomes with an abundance of carbohydrate-active enzyme family isoforms

Sara E. Pidcock
Queen's University Belfast

Timofey Skvortsov
Queen's University Belfast

Stephen J. Courtney
Queen's University Belfast

Christopher J. Creevey
Queen's University Belfast

Sharon A. Huws (✉ S.Huws@qub.ac.uk)
Queen's University Belfast

Short Report

Keywords: rumen, Butyrivibrio, Pseudobutyrivibrio, pangenome, evolution, function, phylogeny, taxonomy, metagenome-assembled genome, MAG

Posted Date: November 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2289353/v1

Abstract

Background

Gut microbiomes are crucial for host nutrition due to their feed energy-harvesting capacity. In the rumen microbiome *Butyrivibrio* and *Pseudobutyrivibrio* dominate and play a key role in harvesting dietary energy. Within these genera, five rumen species have been classified (*Butyrivibrio fibrisolvens*, *B. hungatei*, *B. proteoclasticus*, *Pseudobutyrivibrio ruminis* and *P. xylanivorans*) and more recently an additional sixth *Butyrivibrio* sp. group was added. Nonetheless, in recent years the explosion in available metagenomically assembled genomes (MAGs) offer a new insight into their taxonomy and function. Consequently, in this study we analysed the pangenome and function of 64 MAGs and 71 culture representatives of *Butyrivibrio* and *Pseudobutyrivibrio*.

Results

Using MASH and ANI we demonstrate that the 135 *Butyrivibrio* and *Pseudobutyrivibrio* genomes from MAGs and pure culture cluster into 59 genera and 75 species. Pangenome analysis of 8 groups containing > 4 strains showed average core genome percentages of between 38.49–76.67%. In turn, the accessory genome percentages for the 8 groups were between 23.33% and 65.39%. The most abundant GH families found in the 8 groups were GH13, GH3, GH43, GH2, GH25, GH31, GH36, and GH5 in descending order. Dendograms of the GH families show extensive variation, and expression of 20.45–47.11% of the genes is observed in metatranscriptome datasets.

Conclusions

Our findings demonstrate that Butyrivibrio and *Pseudobutyrivibrio* genomes cluster into 59 genera and 75 species. The 8 groups containing 4 or more genomes that were carried forward possess open genomes with extensive genomic diversity. The MAGs, alongside genomes for cultured isolates, contained an expansive repertoire of glycosyl hydrolase isoforms, which likely facilitate the symbiotic breakdown of plant matter under dietary perturbations allowing a competitive advantage and driving niche specialisation. This study has enabled a substantial enhancement in our understanding of the functional capacity and taxonomy of the dominant rumen isolates, *Butyrivibrio* and *Pseudobutyrivibrio* by utilising all recently published rumen MAGs.

Background

The genus *Butyrivibrio* is composed of bacteria that stain Gram negative, are non-spore forming, strictly anaerobic, with a straight to curved rod morphology. The cells are 0.3–0.8µm wide by 1.0–5.0µm long and can occur singly or in chains. They are flagellated, being monotrichous or lophotrichous, and have polar or sub-polar flagella. They can be motile or non-motile. Bergey's Manual of Systematics of Archaea and Bacteria [1] states that *Butyrivibrio* have a chemo-organotropic metabolism that utilises carbohydrates as a primary substrate group, producing butyrate (and sometimes lactate) as an end product. 16S rRNA gene sequences place *Butyrivibrio* strains with cluster XIVa of the *Clostridium* subphylum; they are in the family *Lachnospiraceae*, order *Clostridiales*, class *Clostridia*, and phylum *Firmicutes*. Willems and Collins (2009) [1] further describe them as being isolated from the rumen, as well as being found in human, rabbit, and horse faeces, and having a GC% content of 36–45%. At the time of their discovery in 1956, Bryant and Small (1956) [2] noted that the variability in the 48 strains determined to belong to *Butyrivibrio* was "considerable" in terms of morphological and physiological characteristics. They stated that this extensive variability would make it difficult to determine what characteristics should logically be used to define natural species specific patterns.

Sometime later, the genus *Pseudobutyrivibrio* (i.e., *Butyrivibrio*-like) was named [3] and described as being similar to *Butyrivibrio* strains, but sufficiently different, with little proteolytic activity compared with *Butyrivibrio* strains, thus warranting the description of a new genus; *P. ruminis*. Meanwhile, a variety of newly isolated bacteria were being assigned to *B. fibrisolvens* on the basis of butyrate production, leading to the expansion and diversification of an already-variegated taxon [4]. Early 16S rDNA phylogenetic analyses of 40 *Butyrivibrio* strains found 12 different rDNA types, each of which was thought to represent at least a separate species, as well as three groups thought to constitute separate genera [5]. At present, the genera *Butyrivibrio* and *Pseudobutyrivibrio* are composed of six species, some of which have been stipulated to contain multiple genera based on DNA-DNA hybridisation [6, 7]. A recent study [7] analysed the genomes of 40 *Butyrivibrio* strains, and 6 *Pseudobutyrivibrio* strains. They found that, based on full-length 16S rDNA sequences, all *Butyrivibrio* strains clustered separately to all *Pseudobutyrivibrio* strains, with the former composing 3 groups and the latter a single one; these groups were confirmed with Functional Genome Distribution (FGD). Average nucleotide identity (ANI) comparing the six *Pseudobutyrivibrio* to the 40 *Butyrivibrio* gave 70–71% sequence identity, indicating the need for separate genera. They [7] go on to state that this boundary between species within each genus is not as easily defined; for example, in each of the three *Butyrivibrio* clusters, some strains formed their own distinct groupings, demonstrating within-species variability. Our own data [8] were similar to previous findings [7], finding extensive pangenomic variation within species of *Butyrivibrio* and *Pseudobutyrivibrio*, with evidence of the existence of 32 genera and 42 species across *Butyrivibrio* and *Pseudobutyrivibrio*.

However, our taxonomic understanding of the *Butyrivibrio* and *Pseudobutyrivibrio* has been elucidated from pure cultures isolated using traditional microbiological methods. Attempts were made to culture new organisms by the modification of culture methods, but these still do not allow the full picture [9]. Given that *Butyrivibrio* and *Pseudobutyrivibrio* are amongst those which can be difficult to culture [10], the introduction of Metagenome Assembled Genomes (MAGs) from the rumen allows for a more complete look at their taxonomy than ever before. MAGs are isolated genomes from metagenomic sequences, allowing a fuller representation of taxonomic diversity to be investigated without the challenges of culture. Consequently, this study investigated taxonomy and function of MAGs, which were not publicly available when our previous analyses were conducted, alongside the data for genomes from pure cultures of *Butyrivibrio* and *Pseudobutyrivibrio*. This allows a significant step-change in our understanding of the functional capacity and taxonomy of the dominant rumen isolates, *Butyrivibrio* and *Pseudobutyrivibrio.*

Methods

Metagenomically Assembled Genomes used in study

A total of 17,818 MAGs were included in this study [11–15]. All published datasets from ruminants were taken where MAGs were publicly available at the time of data collection.

Isolation of Butyrivibrio and Pseudobutyrivibrio MAGs and taxonomic analysis

In order to isolate MAGS pertaining to belong to *Butyrivibrio* and *Pseudobutyrivibrio* each dataset used underwent a Mash v2.0 [16] sequence identity comparison against the initial 71 *Butyrivibrio* and *Pseudobutyrivibrio* genomes (obtained as per [8]), comparing each MAG to each of the 71 genomes. The MAGs were compiled into a reference.msh file, which was subsequently used as a database to be compared against each of the 71 genomes on an individual basis. Mash was run with the default parameter of a k-mer length of 21, but the sketch value was changed from 1000 to 5000 given the divergent nature of the sequences, as this allowed more diversity to be included rather than risk the exclusion of MAGs which may belong to *Butyrivibrio* and *Pseudobutyrivibrio* at this stage. All hits with a Mash value of less than or equal to 0.3 (which approximates to an ANI of 70%, indicating that organisms are within the same family) were retained for further analysis. Mash was used as a faster precursor to ANI as an initial screen due to the number of MAGs included in this study.

In order to then assess Average Nucleotide Identity and ensure that the MAGs chosen belonged to *Butyrivibrio* and *Pseudobutyrivibrio*, the MAGs identified initially using Mash underwent a similar process using FastANI V1.32 [17], with each respective group of MAGs forming the query groups, and the 71 *Butyrivibrio* and *Pseudobutyrivibrio* genomes being the reference sequences, in order to assess the comparative identity of the query sequences to the reference sequences. The "–matrix" parameter was used, creating a phylip-formatted triangular matrix to facilitate analysis of the resulting data, and "–maxmatch" was used to include all Nucmer options. FastANI only provides output with an ANI of approximately 80% sequence identity or higher to the reference sequences. All MAGs still pertaining to *Butyrivibrio* and *Pseudobutyrivibrio* following output by FastANI (i.e. by having greater than ~ 80% sequence identity to a reference sequence) were put forward for analysis via PyANI in order to further resolve taxonomy. PyANI V0.2.11 [18] was run on the resulting 64 MAGs and 71 *Butyrivibrio* and *Pseudobutyrivibrio* genomes with the options "-g" for graphical output, "-f" to force writing into an existing directory, and "-m ANIm" to choose MUMmer as the alignment method.

Annotation and Pangenome analysis of Metagenomically-assembled Butyrivibrio and Pseudobutyrivibrio genomes

The *Butyrivibrio* and *Pseudobutyrivibrio* MAGs were annotated with Prokka V1.12 [19] via the Galaxy [20] platform, V1.14.6. Following this, annotated nucleotide files (.ffn) were analysed in the 8 groups (having > 4 representative genomes based on ANI analysis) via Spine V0.3.1 [21], to find core and accessory percentages. These were obtained for each .ffn file through its respective “coords.txt” file, and a mean taken to find the average core and accessory percentages for each of the 8 groups. The threshold for calling a core gene was that it was present in 100% of genomes, minimum NUCmer sequence identity for an alignment was set to 85%, minimum core genome segment size to output as 10bp, and maximum distance between core genome segments is 10bp. Functional annotation of the pangenome (split into core and accessory) was achieved using EggNOG mapper V2.1.7 [22]. Sequences were uploaded in CDS format, and all parameters were default. The COG categories were visualised as a stacked histogram. Glycosyl hydrolase (GH) families were also identified for these 8 groups using the dbCAN2 meta-server [23], selecting to run HMMer with output values having an E-Value < 1e-15 and coverage > 0.35.

Carbohydrate Digestion Capability

Glycosyl hydrolase (GH) families were identified using the dbCAN2 meta-server [23], selecting to run HMMer with output values having an E-Value < 1e-15 and coverage > 0.35 for all *Butyrivibrio* and *Pseudobutyrivibrio* MAG and pure culture genomes. From the output, only GH families were selected to be visualised using a stacked histogram. Only GH families with more than 50 total instances were included to facilitate visualisation. GH sequences were extracted using Samtools V1.9 [24], aligned using MUSCLE V3.8.31 [25] using default parameters, with each alignment containing an outlier sequence from the respective GH family. To determine which genes were actively expressed and therefore were likely real and not false genes, Bowtie2 was used to align reads from 13 metatranscriptome samples [26] (SRR IDs: SRR1138694, SRR1138697, SRR1206249, SRR873450, SRR873451, SRR873452, SRR873453, SRR873455, SRR873456, SRR873459, SRR873460, SRR873461, SRR873465) to nucleotide sequences of 2,425 GH genes. The ‘–very-fast –no-unal -p 1- –reorder” options were used to carry out end-to-end alignments of the reads to the genes. Subsequently, phylogenetic tree files were created using IQ-TREE V1.6.1 [27]. These were then visualised using iToL [28], and expression data and sequence origin displayed using the iToL template sheets.

Results

Isolation and taxonomy of Metagenomically-assembled Butyrivibrio and Pseudobutyrivibrio genomes

From the initial 5 datasets containing a total of 17,818 MAGs, a total of 2,368 MAGs were returned by Mash as being related on a family level (Family Lachnospiraceae; Supplementary Excel 3.1). Of these, 64 were returned by FastANI, meaning that they were higher than approximately 80% sequence identity to *Butyrivibrio* and *Pseudobutyrivibrio* (Supplementary Excel 3.2). These 64 were combined with the original 71 *Butyrivibrio* and *Pseudobutyrivibrio* genomes, giving 135 genomes in the more resolved PyANI analysis. The resulting sequences, when analysed by PyANI, showed a total of 75 potential species in 59 potential genera (Fig. 1A &B).

Pangenomics of Metagenomically-assembled Butyrivibrio and Pseudobutyrivibrio genomes

Alongside this taxonomic data, it was shown that eight groups had 4 or more strains within them (7, 10, 4, 5, 4, 4, 4, and 5 genomes respectively; Fig. 1A). Coverage was low outside of these groups, indicating a multitude of genera (Fig. 1B). Pangenomic analysis showed that the 8 groups had average core genome percentages of 38.49%, 43.24%, 42.15%, 44.78%, 34.61%, 50.20%, 58.44%, and 76.67% respectively. In turn, the accessory genome percentages for the 8 groups were as follows: 61.51%, 56.76%, 57.85%, 55.22%, 65.39%, 49.80%, 41.66%, and 23.33%.

Functional annotation of Metagenomically-assembled Butyrivibrio and Pseudobutyrivibrio genomes

The eight identified groups underwent functional annotation by EggNOG (Fig. 2) which showed that approximately 20–30% of each genome was dedicated to unknown functions. Aside from this, the function with the most genes attributed to it was carbohydrate transport and metabolism, for both the core and accessory genome, although the latter appears to have a slightly smaller proportion of genes dedicated to this.

The most abundant GH families found in the 8 groups were GH13, GH3, GH43, GH2, GH25, GH31, GH36, and GH5. Within the 8 groups, a total of 474 GH13's were identified, 316 GH3's, 248 GH43's, 242 GH2's, 172 GH25's, 104 GH31's, 95 GH36's, and 88 GH5's (Fig. 3). Whilst variation can be seen between groups, they appear to have similar proportions of glycosyl hydrolase families overall. Dendograms for GH family 2, 3, 5, 13, and 43 were produced, and expression data [26] showed that, for the GH2 tree, 114 genes were found to be expressed to some extent in at least one metatranscriptome dataset. GH3 147 genes, for GH5 18 genes, GH13 207 genes, and GH43 76. This equates to 47.11%, 46.52%, 20.45%, 43.67%, and 30.65% of the total genes in that GH family respectively (Figs. 4–8).

Discussion

The rumen is a highly divergent habitat, being home to an estimated 7400 species of bacteria and 1400 archaea, as well as various fungi and viruses [29]. *Butyrivibrio* and *Pseudobutyrivibrio* (Family *Lachnospiraceae*) are thought to make up a considerable proportion of this, reflecting approximately 4.8% and 2.3% of all *Clostridia* sequences respectively. *Lachnospiraceae* account for 23.8% of *Clostridia* sequences [29]. In this study we analysed the taxonomy of *Butyrivibrio* and *Pseudobutyrivibrio* to a level not previously possible, through the addition of recently available MAG data. We found in comparison that 13.29% of our 17,818 MAG samples were related to available pure culture genomes of *Butyrivibrio* and *Pseudobutyrivibrio*. Essentially 64 MAGs had greater than 80% sequence identity (as determined by FastANI) to the cultured 71 strains of *Butyrivibrio* and *Pseudobutyrivibrio*, and when analysed via PyANI these cumulative 135 sequences (64 MAGs and 71 pure culture genomes) had 95% identity thresholds that indicate that 75 species likely exist. The coverage, in turn, suggests 59 potential genera. The 8 groups formed (which contained > 4 strains) are similar to previously determined groups [8], with the largest group in both studies being formed by *B. fibrisolvens* strains. Whilst the previous *B. fibrisolvens* group contained 8 genomes, the new group adds two MAGs; 557N_min2k_bin.118 [14], and RGIG578 [15].

Pangenomic analysis of the 8 groups revealed core genome percentages of between 34.61% (group 5) to 76.67% (group 8). Generally, as more genomes are added to a population, the core genome gets smaller and the accessory genome larger [30]. This is more likely to be the case in “open” pangenomes (whereby each new addition is likely to add an abundance of accessory genes) than in “closed” pangenomes (in which new genomes will introduce very few new genes). Given that *Butyrivibrio* are considered to have an open pangnome [31], our finding that they have a smaller core genome than, for example, *Staphylococcus lugdunensis*, whose core genes represent 86–88% of the entire genome [21] corroborates our hypothesis that these genera have very open genomes. An open pangenome is proposed to be typical of organisms that colonise multiple environments, and are more likely to exchange genetic information [32]; given that *Butyrivibrio* have been isolated from ruminants worldwide, regardless of geography, host animal, host feed, etc. [33], it is logical that they require an open pangenome in order to facilitate metabolic function in a diverse range of habitats. This is supported by previous findings [34], which suggest that the ability to migrate to new niches is one of the biggest determining factors in pangenome size.

The majority of all genes annotated by EggNOG for the 8 groups were of unknown function, as is typical, with 40–60% of newly identified genes not being assigned a known function [35]. Of the remaining genes, most were annotated as being involved with typically core functions, such as carbohydrate transport and metabolism, amino acid transport and metabolism, cell wall, membrane, and envelope biogenesis etc. Carbohydrate metabolising and various transport genes have been identified as being part of the core genome of *Butyirvibrio* before [8]. This is likely due to the fact that *Butyirivbrio* and *Pseudobutyrivibrio* are capable of degrading a wide range of carbohydrates, necessitating a range of enzymes to facilitate this [36]. Little variation in proportion of functions between the core and accessory genomes was seen; whilst this may initially appear surprising, the functional categories are vague, indicating that both the core and accessory genomes contain genes that orchestrate these functions. This data also suggests that these are not the same genes, and will catalyse different reactions within that functional bracket, i.e. even if the enzymes are within the same functional group, they may have varying specificities.

Each of the 8 groups formed by ANIm analysis contained a rich diversity of glycosyl hydrolases. Of all glycosyl hydrolase families, GH13 was the most abundant. This is consistent with the findings of Neves *et al*. (2021) [37], who found GH13 to be the most abundant of 61 families of glycosyl hydrolase identified in a rumen metatranscriptome. Members of GH13 are primarily amylases, but can also degrade a broad spectrum of starches [38]. GH13s have been shown to be significantly up-regulated in *B. hungatei* MB2003 when grown on pectin [36] and ruminant feed can range from 3–25% pectin content [39]. Therefore these GH13’s play a key role in feed degradation, which explains their abundance. GH family 3 enzymes were also present in high abundance; GH3 enzymes are primarily β-glucosidases, which are capable of hydrolysing the β-glucosidic linkages in complex carbohydrate molecules. GH3 enzymes are mostly intracellular [40], and have also been found to be abundantly present in newly isolated *Butyrivibrio* strains CB08, XB500-5, and X503, which it is suggested facilitates their high levels of cellobiohydrolase and xylobiohydrolase activity [31]. Alongside endo- and exoglucanases, β-glucosidases play a key role in the metabolism of cellulose, the most abundant organic biopolymer in the biosphere; endo- and exoglucanases act synergistically to hydrolyse glucosidic bonds to generate various oligosaccharides.

These oligosaccharides are hydrolysed in turn, releasing glucose, cellobiose and shorter oligosaccharides; β-glucosidase then breaks down the cellobiose and short oligosaccharides into glucose [41]. Given that the primary constituent of plant fibre is cellulose, it is imperative that rumen microorganisms (and ruminants in turn) are able to efficiently catalyse its breakdown in order to improve metabolic efficiency. The most abundant endoglucanases were GH5; endoglucanases help catalyse the breakdown of common cattle feed such as corn silage, ultimately improving cattle productivity [42].

The GH dendograms show extensive sequence diversity in all GH families analysed. In general, proteins annotated as being of the same type (e.g. cyclomaltodextrinases) grouped together, regardless of source (i.e. MAG or Hungate), although same-source sequences tended to form sub-groups within these. It is thought that multiple enzyme isoforms allow a competitive advantage in fluctuating environmental conditions, with isoform diversity playing an important role in maintaining microbial niche specialisation [43]. As such, the incredible GH diversity revealed in these sequences is likely to be evolutionarily advantageous, providing the ruminant with continuous nutrition when faced with dietary perturbations. The expression of such a high proportion of these GH's shows that a multitude of dietary enzymes are being expressed at any one time; this reveals the complex nature of the rumen, and the potential metabolic processes that it hosts, with a range of organisms contributing to different aspects of feed metabolism simultaneously [44, 45]. It should be noted that this GH diversity, being derived from MAG sequences, are only predictions and would require further analysis to determine whether they are an accurate reflection of natural diversity.

Conclusions

This study allowed a significant step-change in our understanding of the functional capacity and taxonomy of the dominant rumen isolates, *Butyrivibrio* and *Pseudobutyrivibrio.* We focused on pangenomic and functional analysis of 64 MAGs and 71 culture representatives of *Butyrivibrio* and *Pseudobutyrivibrio*. Our study suggests the existence of 59 genera and 75 species, far beyond what was previously understood. Not only do our findings demonstrate extensive genomic diversity, but we also show an expansive repertoire of glycosyl hydrolase isoforms (many of which were found to be expressed), which are thought to facilitate the symbiotic breakdown of plant matter. Despite extensive genomic variation, all strains appear to fulfil a similar functional role, supporting the hypothesis that this heterogeneity lies within degradative enzyme isoforms, allowing a competitive advantage and driving niche specialisation.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets generated and/or analysed during the current study are available in Zenodo: 10.5281/zenodo.7335268

Competing interests

The authors declare that they have no competing interests

Funding

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).

Authors' contributions

S.A.H. and S.E.P. conceptualised the research and led the project. T.S., S.J.C, and C.J.C helped S.E.P. with computational analysis and discussions regarding project direction. S.E.P. and S.A.H. drafted the manuscript.

Acknowledgements

Not applicable

References

1. Willems AC, M. Butyrivibrio. W.B W, editor. New York, USA: John Wiley & Sons; 2009. 1–20 p.
2. Bryant MP, Small N. Characteristics of two new genera of anaerobic curved rods isolated from the rumen of cattle. J Bacteriol. 1956;72(1):22–6.
3. Van Gylswyk NH, H.; Rainey, F. *Pseudobutyrivibrio ruminis* gen. nov., sp. nov., a Butyrate-Producing Bacterium from the Rumen That Closely Resembles *Butyrivibrio fibrisolvens* in Phenotype. *International Journal of Systematic Bacteriology*. 1996;2(46):559 – 63.
4. 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(Pt 1):201–9.
5. Willems A, Amat-Marco M, Collins MD. Phylogenetic analysis of Butyrivibrio strains reveals three distinct groups of species within the Clostridium subphylum of the gram-positive bacteria. Int J Syst Bacteriol. 1996;46(1):195–9.
6. Mannarelli B. Deoxyribonucleic Acid Relatedness among Strains of the Species Butyrivibrio fibrisolvens. International Journal of Systematic Bacteriology. 1988;4(38):340–7.

7. Palevich N, Kelly WJ, Leahy SC, Denman S, Altermann E, Rakonjac J, et al. Comparative Genomics of Rumen Butyrivibrio spp. Uncovers a Continuum of Polysaccharide-Degrading Capabilities. Appl Environ Microbiol. 2019;86(1).
8. Pidcock SE, Skvortsov T, Santos FG, Courtney SJ, Sui-Ting K, Creevey CJ, et al. Phylogenetic systematics of Butyrivibrio and Pseudobutyrivibrio genomes illustrate vast taxonomic diversity, open genomes and an abundance of carbohydrate-active enzyme family isoforms. Microb Genom. 2021;7(10).
9. Nyonyo T, Shinkai T, Mitsumori M. Improved culturability of cellulolytic rumen bacteria and phylogenetic diversity of culturable cellulolytic and xylanolytic bacteria newly isolated from the bovine rumen. FEMS Microbiol Ecol. 2014;88(3):528–37.
10. Bryant MP, Robinson IM. Some nutritional characteristics of predominant culturable ruminal bacteria. J Bacteriol. 1962;84:605–14.
11. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun. 2018;9(1):870.
12. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37(8):953–61.
13. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, et al. 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. Genome Biol. 2020;21(1):229.
14. Glendinning L, Genc B, Wallace RJ, Watson M. Metagenomic analysis of the cow, sheep, reindeer and red deer rumen. Sci Rep. 2021;11(1):1990.
15. Xie F, Jin W, Si H, Yuan Y, Tao Y, Liu J, et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. Microbiome. 2021;9(1):137.
16. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132.
17. 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(1):5114.
18. Pritchard LG, R.; Humphris, S.; Elphinstone, J; Toth, I. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Analytical Methods. 2016;1(8):12–24.
19. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
20. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537-W44.
21. 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.

22. Huerta-Cepas JS, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S.; Cook, H.; Mende, D.; Letunic, I.; Rattei, T.; Jensen, L.; von Mering, C; Bork, P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2018;1:309–14.
23. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95-W101.
24. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).
25. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
26. Shi W, Moon CD, Leahy SC, Kang D, Froula J, Kittelmann S, et al. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res. 2014;24(9):1517–25.
27. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
28. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.
29. Kim M, Morrison M, Yu Z. Status of the phylogenetic diversity census of ruminal microbiomes. FEMS Microbiol Ecol. 2011;76(1):49–63.
30. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol. 2010;60(4):708–20.
31. Sengupta K, Hivarkar SS, Palevich N, Chaudhary PP, Dhakephalkar PK, Dagar SS. Genomic architecture of three newly isolated unclassified Butyrivibrio species elucidate their potential role in the rumen ecosystem. Genomics. 2022;114(2):110281.
32. Medini DD, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The Microbial Pan-genome. Current Opinion in Genetics Development. 2005;6(15):589–94.
33. Henderson GC, F.; Ganesh, S.; Jonker, A.; Young, W.; Janssen, P. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Scientific Reports. 2015;1.
34. McInerney JM, A.; O'Connell, M. Why prokaryotes have pangenomes. Nature Microbiology. 2017;2(4).
35. Vanni C, Schechter MS, Acinas SG, Barberan A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. Elife. 2022;11.
36. Palevich N, Kelly WJ, Ganesh S, Rakonjac J, Attwood GT. Butyrivibrio hungatei MB2003 Competes Effectively for Soluble Sugars Released by Butyrivibrio proteoclasticus B316(T) during Growth on Xylan or Pectin. Appl Environ Microbiol. 2019;85(3).
37. Neves ALA, Yu J, Suzuki Y, Baez-Magana M, Arutyunova E, O'Hara E, et al. Accelerated discovery of novel glycoside hydrolases using targeted functional profiling and selective pressure on the rumen microbiome. Microbiome. 2021;9(1):229.

38. Labes A, Karlsson EN, Fridjonsson OH, Turner P, Hreggvidson GO, Kristjansson JK, et al. Novel members of glycoside hydrolase family 13 derived from environmental DNA. Appl Environ Microbiol. 2008;74(6):1914–21.
39. Marounek MD, D. Metabolism of pectin in rumen bacteria Butyrivibrio fibrisolvens and Prevotella ruminicola. Letters in Applied Microbiology. 1999;6(29):429–33.
40. Kelly WJ, Leahy SC, Altermann E, Yeoman CJ, Dunne JC, Kong Z, et al. The glycobiome of the rumen bacterium Butyrivibrio proteoclasticus B316(T) highlights adaptation to a polysaccharide-rich environment. PLoS One. 2010;5(8):e11942.
41. Oladoja EO, O.; Adamu, B.; Balogun, A.; Musa, O. Microbial β-glucosidase: Source, production and applications. Journal of Applied & Environmental Microbiology. 2017;1(1):14–22.
42. Eun JB, K. Assessment of the potential of feed enzyme additives to enhance utilization of corn silage fibre by ruminants. Canadian Journal of Animal Science. 2008;1(88):97–106.
43. Rubino F, Carberry C, Waters SM, Kenny D, McCabe MS, Creevey CJ. Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. ISME J. 2017;11(6):1510.
44. Castillo C, Hernandez J. Ruminal Fistulation and Cannulation: A Necessary Procedure for the Advancement of Biotechnological Research in Ruminants. Animals (Basel). 2021;11(7).
45. Mizrahi I, Wallace RJ, Morais S. The rumen microbiome: balancing food security and environmental impacts. Nat Rev Microbiol. 2021;19(9):553–66.

Figures

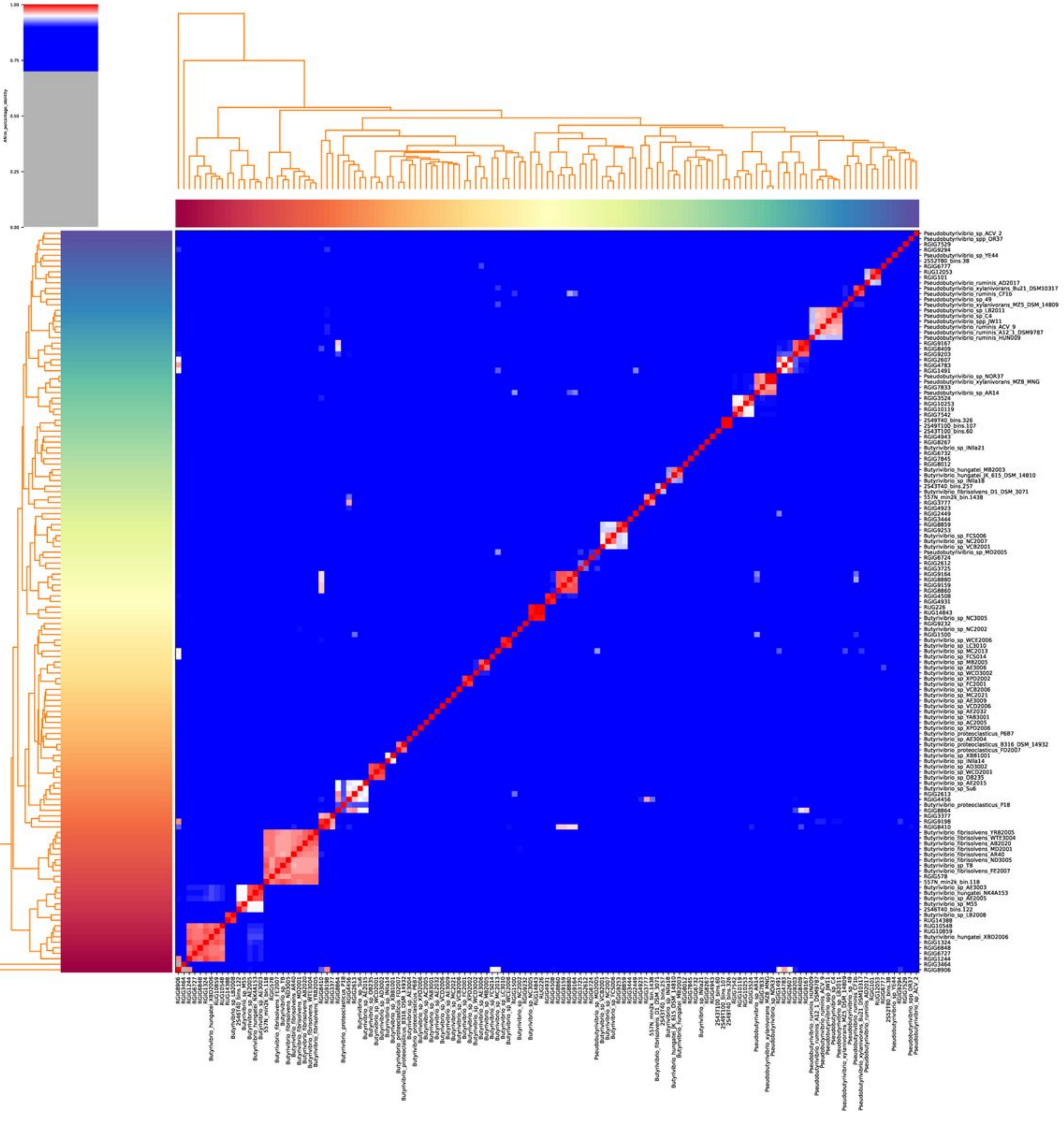

Figure 1

A and B - ANI comparison of 71 strains of *Butyrivibrio* and *Pseudobutyrivibrio* and 64 metagenome-assembled genomes using pyani.py (https://github.com/widdowquinn/pyani/tree/version_0_2, [18]) with the MUMmer alignment option (a) (cells in the heat map that are coloured red have >95% sequence identity, whilst blue cells have <95% sequence identity, 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% coverage, and as coverage reaches 50% the cells are coloured white).

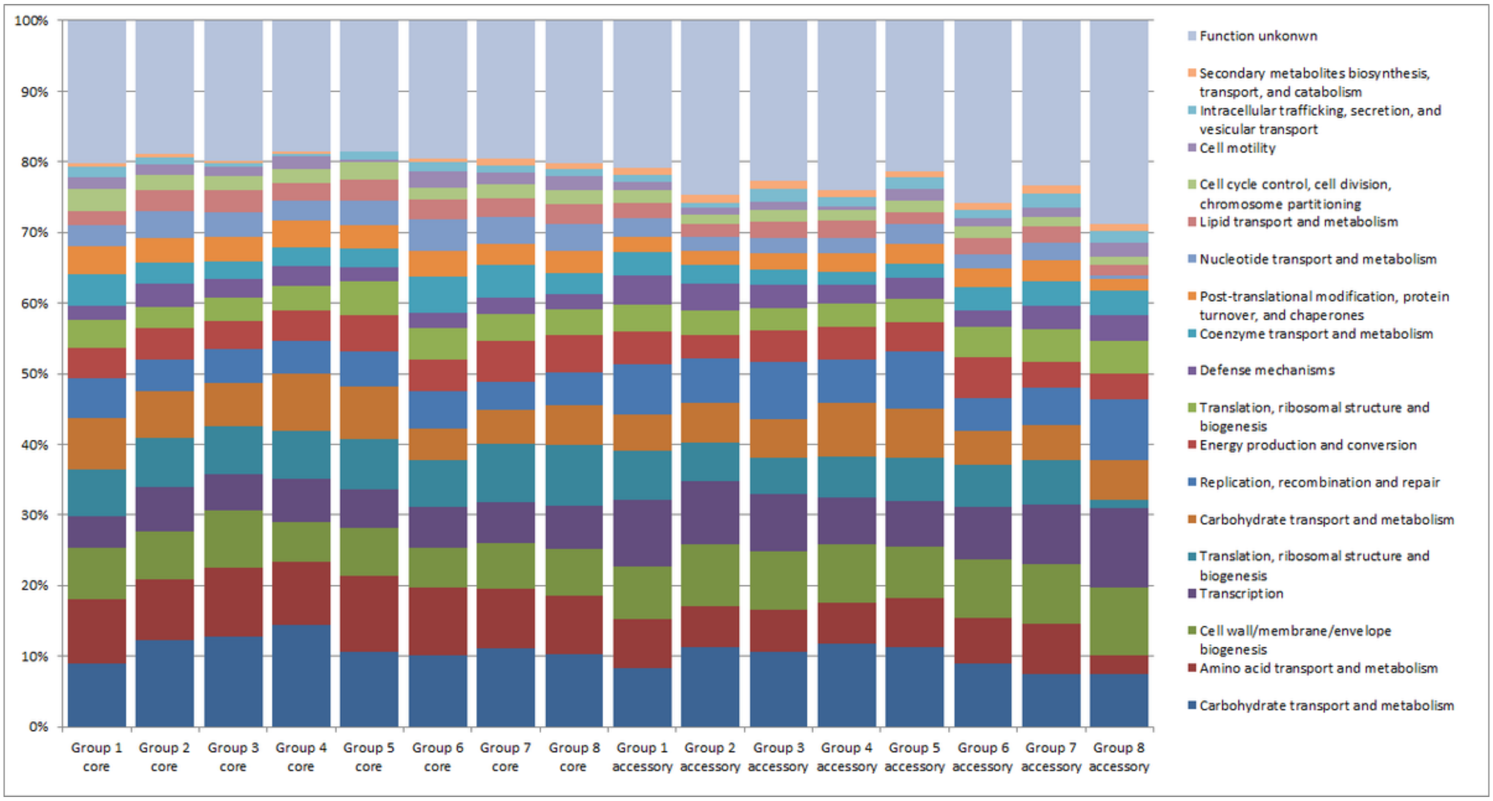

Figure 2

Functional annotation [22] of 8 groups of *Butyrivibrio* and *Pseudobutyrivibrio*strains, split into core and accessory by Spine V0.3.1 [21].

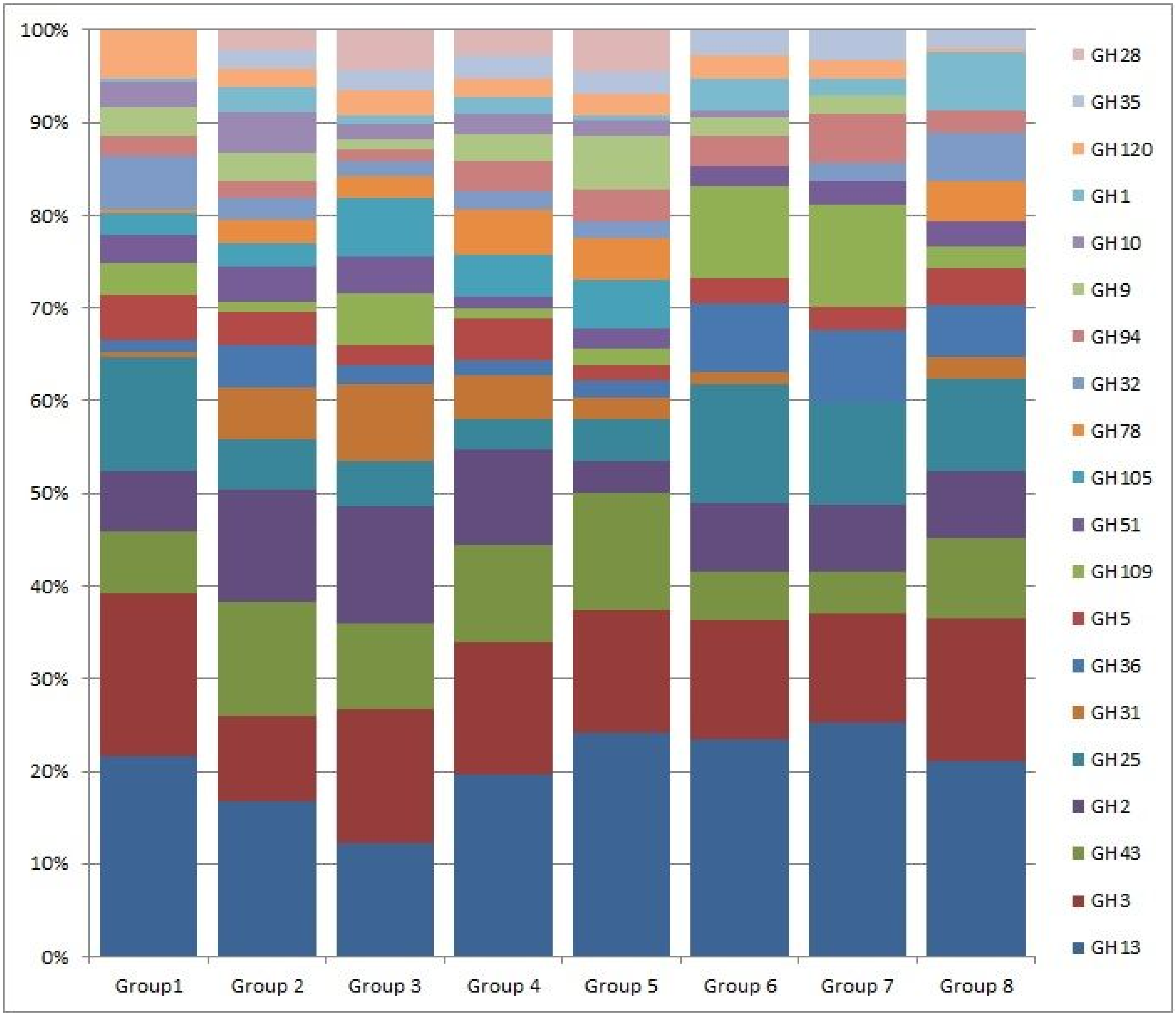

Figure 3

proportions of glycosyl hydrolase (GH) families found in groups of *Butyrivibrio* and *Pseudobutyrivibrio* strains and MAGs. The GHs were identified via the dbCAN metaserver (https://bcb.unl.edu/dbCAN2/blast.php, [23]), and the groups determined via PyANI (https://github.com/widdowquinn/pyani/tree/version_0_2, [18]).

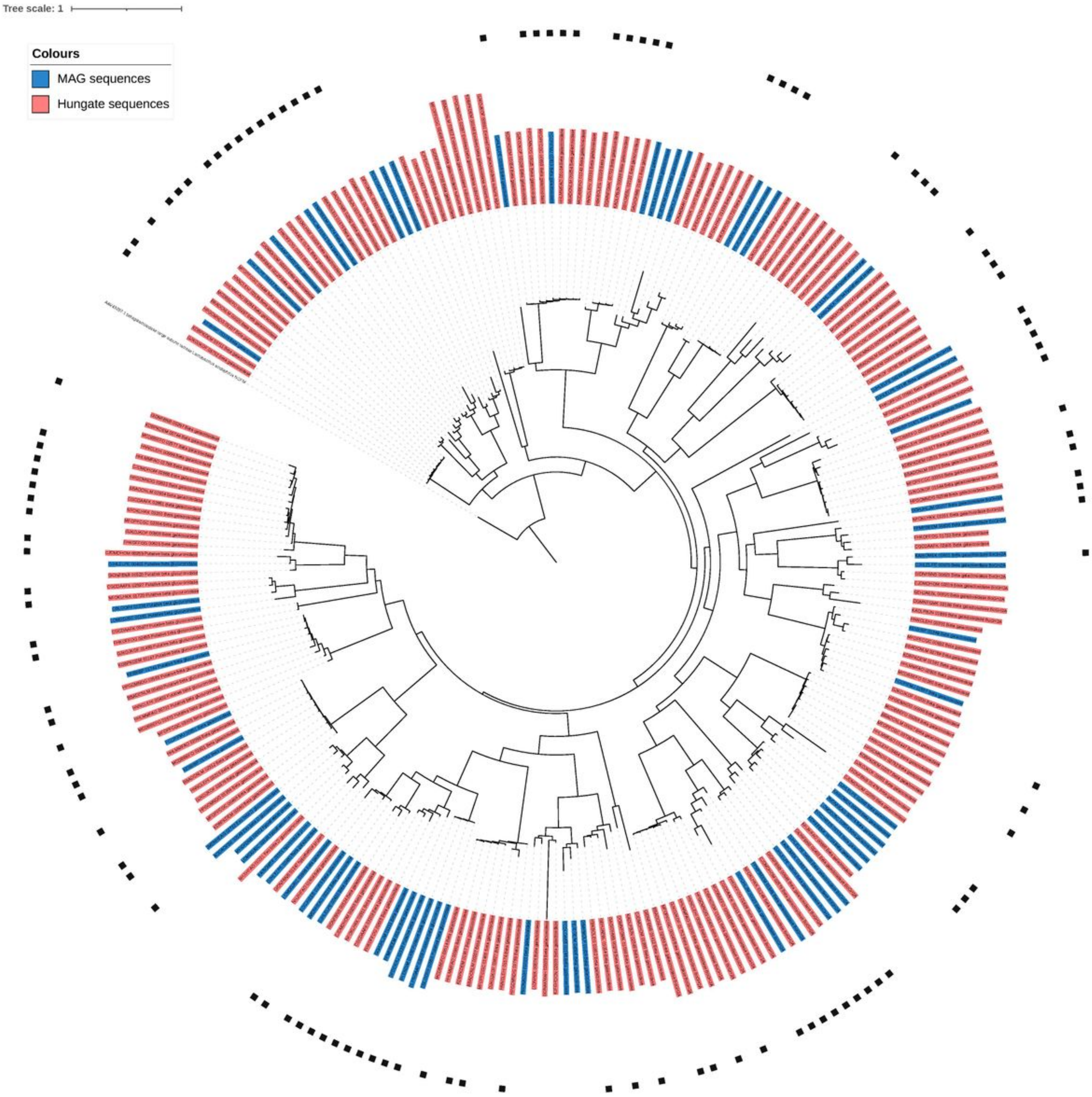

Figure 4

dendogram showing the relatedness of all Glycosyl Hydrolase family 2 genes found in 71 culture-derived genomes and 64 MAG-derived genomes. Colours denote the origin of the sequence, with blue indicating a MAG sequence and red a genome from the Hungate collection. The presence of a black square on the outermost layer indicates that that gene was found to be present in a metatranscriptome dataset [26]. The tree is rooted using a β-galactosidase large subunit sequence from *Lactobacillus acidophilus* NCFM.

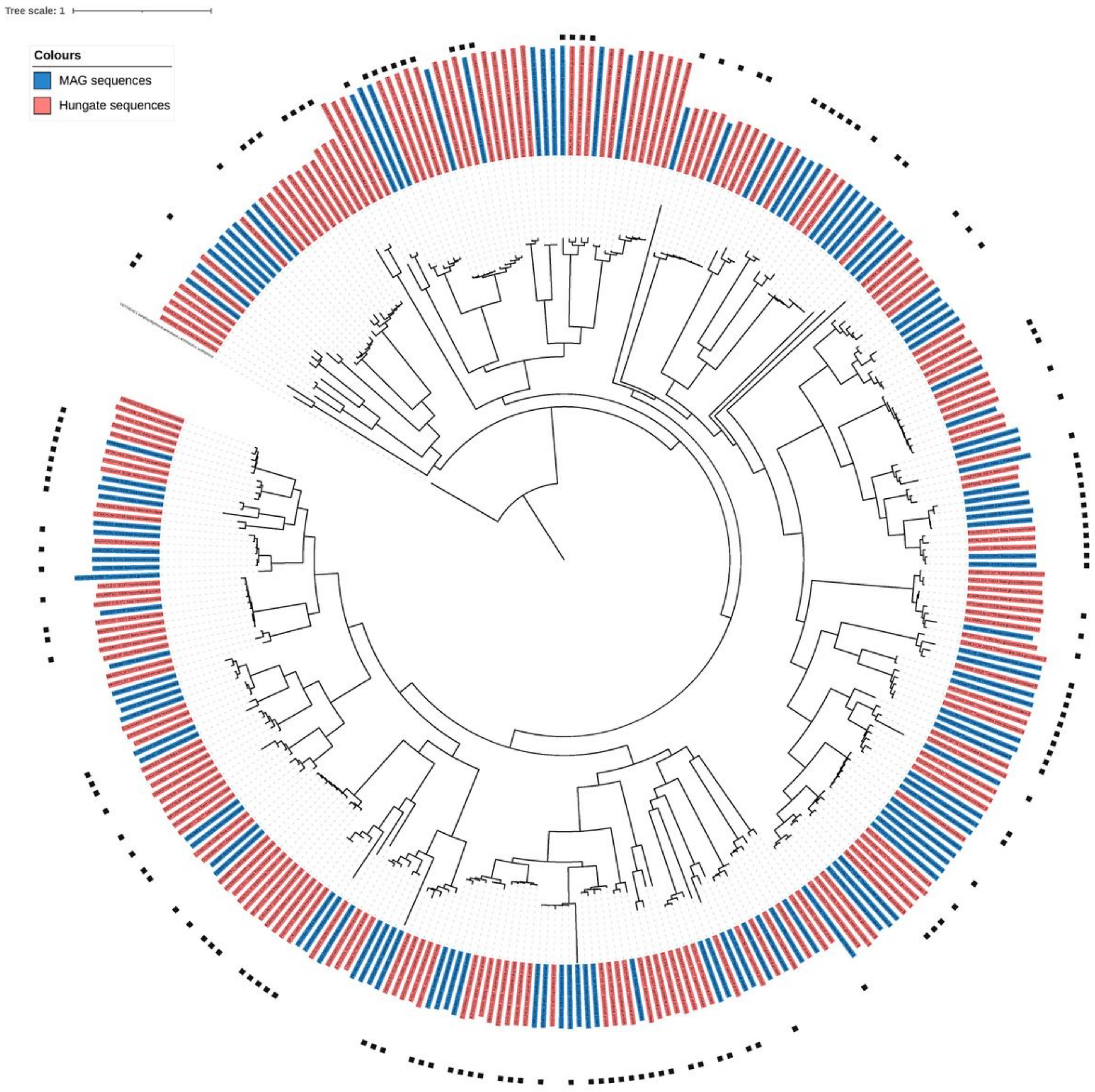

Figure 5

dendogram showing the relatedness of all Glycosyl Hydrolase family 3 genes found in 71 culture-derived genomes and 64 MAG-derived genomes. Colours denote the origin of the sequence, with blue indicating a MAG sequence and red a genome from the Hungate collection. The presence of a black square on the outermost layer indicates that that gene was found to be present in a metatranscriptome dataset [26]. The tree is rooted using a β-N-acetylhexosaminidase sequence from *Lactobacillus acidophilus* NCTC13720.

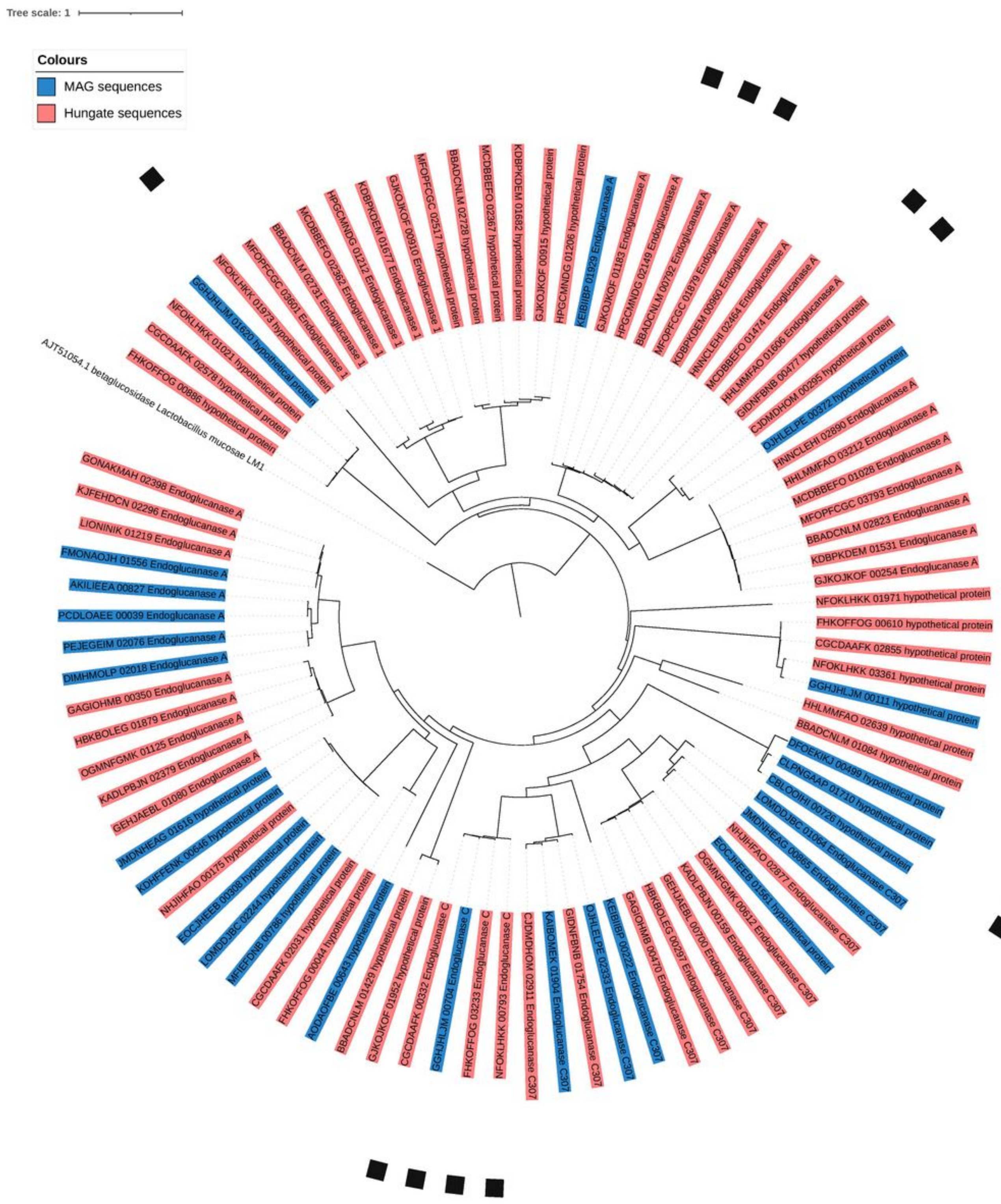
Tree scale: 1
Colours
MAG sequences
Hungate sequences
AJT51054.1 betaglucosidase Lactobacillus mucosae LM1
FHKOFFOG 00886 hypothetical protein
CGCDAAFK 02578 hypothetical protein
NFOKLHKK 01021 hypothetical protein
GGHJHLJM 01620 hypothetical protein
NFOKLHKK 01973 hypothetical protein
MFOPFCGC 03601 Endoglucanase 1
BBADCNLM 02731 Endoglucanase 1
MCDBBEFO 02362 Endoglucanase 1
HPGCMNDG 01212 Endoglucanase 1
KDBPKDEM 01677 Endoglucanase 1
GJKOJKOF 00910 Endoglucanase 1
MFOPFCGC 02517 hypothetical protein
BBADCNLM 02728 hypothetical protein
MCDBBEFO 02367 hypothetical protein
KDBPKDEM 01692 hypothetical protein
GJKOJKOF 00915 hypothetical protein
HPGCMNDG 01206 hypothetical protein
KEIBIIBP 01929 Endoglucanase A
GJKOJKOF 01183 Endoglucanase A
HPGCMNDG 02149 Endoglucanase A
BBADCNLM 00792 Endoglucanase A
MFOPFCGC 01879 Endoglucanase A
KDBPKDEM 00960 Endoglucanase A
HNNCLEHI 02464 Endoglucanase A
MCDBBEFO 01474 Endoglucanase A
HHLMMFAO 01606 Endoglucanase A
GIDNFBNB 00477 hypothetical protein
CJDMDHOM 00295 hypothetical protein
OJHLELPE 00372 hypothetical protein
HNNCLEHI 02890 Endoglucanase A
HHLMMFAO 03212 Endoglucanase A
MCDBBEFO 01028 Endoglucanase A
MFOPFCGC 03793 Endoglucanase A
BBADCNLM 02823 Endoglucanase A
KDBPKDEM 01531 Endoglucanase A
GJKOJKOF 00254 Endoglucanase A
NFOKLHKK 01971 hypothetical protein
FHKOFFOG 00610 hypothetical protein
CGCDAAFK 02855 hypothetical protein
NFOKLHKK 03361 hypothetical protein
GGHJHLJM 00111 hypothetical protein
HHLMMFAO 02639 hypothetical protein
BBADCNLM 01084 hypothetical protein
DFOEKIKJ 00499 hypothetical protein
CLPNGAAP 01710 hypothetical protein
CBLOOIHI 00726 hypothetical protein
LOMDDJBC 01064 Endoglucanase C307
JMDNHEAG 00885 Endoglucanase C307
NHJIHFAO 02877 Endoglucanase C307
EOCJHEEB 01561 hypothetical protein
OGMNFGMK 00612 Endoglucanase C307
KADLPBJN 00159 Endoglucanase C307
GEHJAEBL 00700 Endoglucanase C307
HBKBOLEG 00297 Endoglucanase C307
GAGIOHMB 00470 Endoglucanase C307
KEIBIIBP 00222 Endoglucanase C307
OJHLELPE 02333 Endoglucanase C307
GIDNFBNB 01754 Endoglucanase C307
AIBOMEK 01904 Endoglucanase C307
CJDMDHOM 02911 Endoglucanase C307
NFOKLHKK 00793 Endoglucanase C
FHKOFFOG 03233 Endoglucanase C
GGHJHLJM 00704 Endoglucanase C
CGCDAAFK 00332 Endoglucanase C
GJKOJKOF 01952 hypothetical protein
BBADCNLM 01429 hypothetical protein
AODAOFBE 00643 hypothetical protein
FHKOFFOG 00044 hypothetical protein
CGCDAAFK 02031 hypothetical protein
MFIEFDNE 00786 hypothetical protein
LOMDDJBC 02244 hypothetical protein
EOCJHEEB 00308 hypothetical protein
NHJIHFAO 00175 hypothetical protein
KOHFFENK 00646 hypothetical protein
JMDNHEAG 01616 hypothetical protein
GEHJAEBL 01080 Endoglucanase A
KADLPBJN 02379 Endoglucanase A
OGMNFGMK 01125 Endoglucanase A
HBKBOLEG 01879 Endoglucanase A
GAGIOHMB 00350 Endoglucanase A
DIMHMOLP 02018 Endoglucanase A
PEJEGEIM 02076 Endoglucanase A
PCDLOAEE 00039 Endoglucanase A
AKILIEEA 00827 Endoglucanase A
FMONAOJH 01556 Endoglucanase A
LIONINIK 01219 Endoglucanase A
KJFEHDCN 02296 Endoglucanase A
GONAKMAH 02398 Endoglucanase A

Figure 6

dendogram showing the relatedness of all Glycosyl Hydrolase family 5 genes found in 71 culture-derived genomes and 64 MAG-derived genomes. Colours denote the origin of the sequence, with blue indicating a MAG sequence and red a genome from the Hungate collection. The presence of a black square on the outermost layer indicates that that gene was found to be present in a metatranscriptome dataset [26]. The tree is rooted using a β-glucosidase sequence from *Lactobacillus mucosae* LM1.

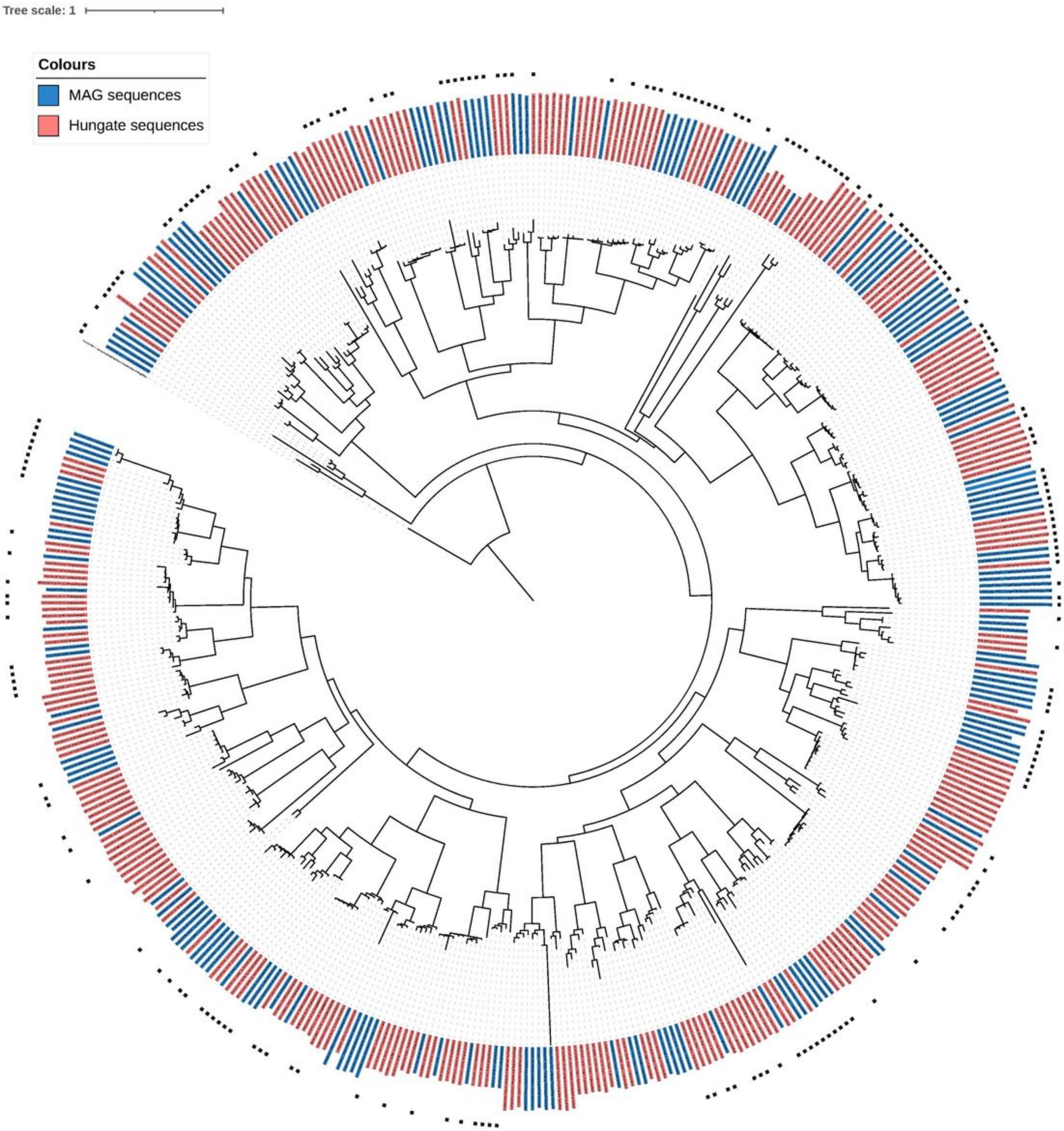

Figure 7

dendogram showing the relatedness of all Glycosyl Hydrolase family 13 genes found in 71 culture-derived genomes and 64 MAG-derived genomes. Colours denote the origin of the sequence, with blue indicating a MAG sequence and red a genome from the Hungate collection. The presence of a black square on the outermost layer indicates that that gene was found to be present in a metatranscriptome dataset [26]. The tree is rooted using a sucrose phosphorylase sequence from *Lactobacillus acidophilus* NCFM.

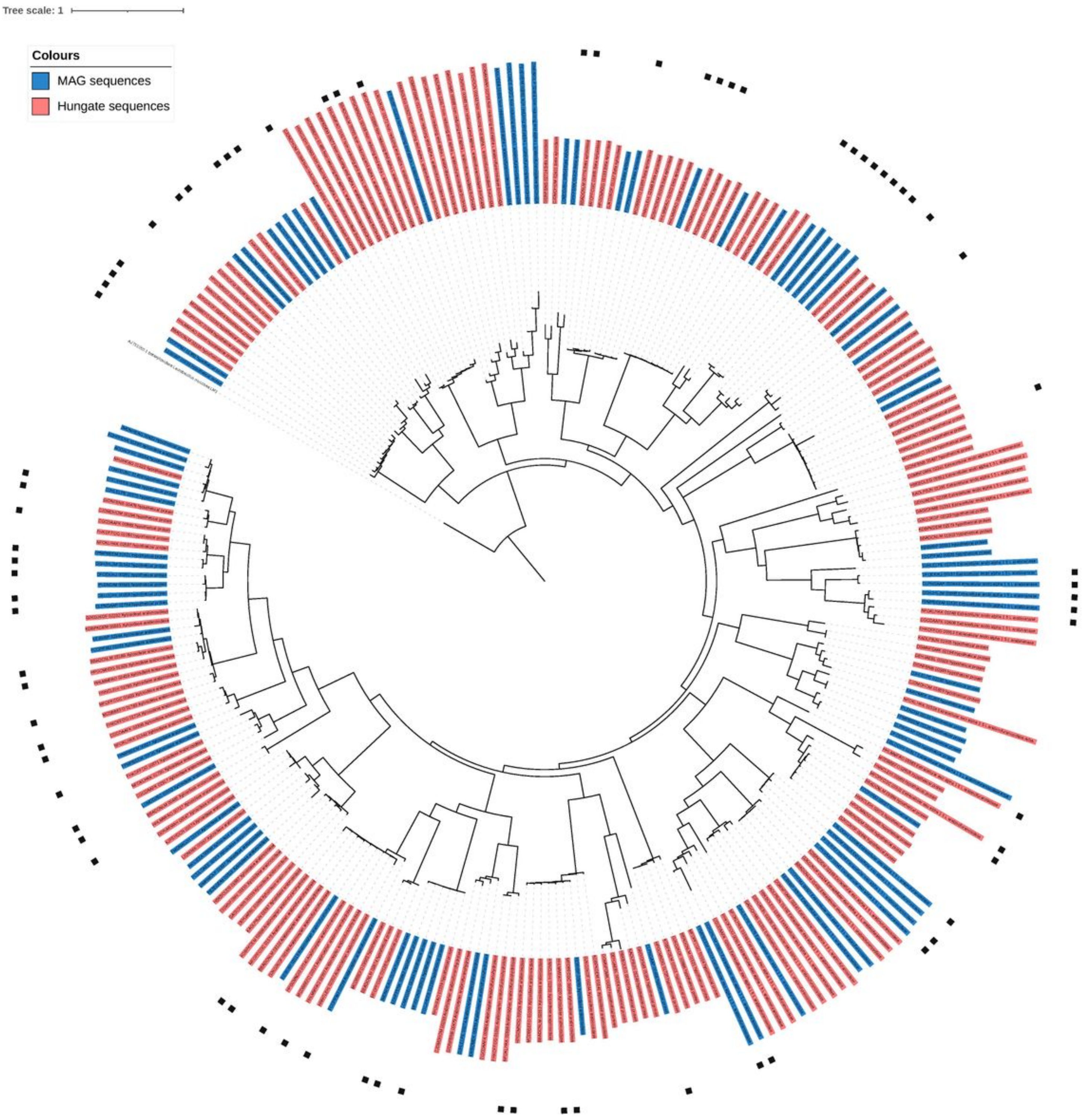

Figure 8

dendogram showing the relatedness of all Glycosyl Hydrolase family 43 genes found in 71 culture-derived genomes and 64 MAG-derived genomes. Colours denote the origin of the sequence, with blue indicating a MAG sequence and red a genome from the Hungate collection. The presence of a black square on the outermost layer indicates that that gene was found to be present in a metatranscriptome dataset [26]. The tree is rooted using a β-xylosidase sequence from *Lactobacillus mucosae* LM1.

Supplementary Files

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

- SupplementaryExcel1Mash.ods
- SupplementaryExcel2FastANI.ods