Gut Microbiome of the Amazon Master of the Grasses Harbors Unprecedented Enzymatic Strategies for Plant Glycans Breakdown

Lucelia Cabral
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Gabriela F Persinoti
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Douglas A Paixao
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Marcele P Martins
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Mariana Chinaglia
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Mariane N Domingues
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Mariana AB Morais
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Mauricio L Sforca
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Renan A. S. Pirolla
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Wesley C Generoso
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Clelton A Santos
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Lucas F Maciel
CNPEM: Centro Nacional de Pesquisa em Energia e Materiais
Nicolas terrapon
AFMB: Architecture et Fonction des Macromolecules Biologiques
Vincent Lombard
AFMB: Architecture et Fonction des Macromolecules Biologiques
Bernard Henrissat
AFMB: Architecture et Fonction des Macromolecules Biologiques
Mario Murakami (mario.t.murakami@gmail.com)

LNBR-CNPEM https://orcid.org/0000-0002-0405-8010

Research

Keywords: capybara, gut microbiome, multi-omics, CAZymes, dietary polysaccharides, plant biomass digestion

Posted Date: November 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-104566/v1

Abstract

Background

Plant biomass is a promising feedstock to replace fossil-based products including fuels, chemicals and materials. However, the high resistance of plant biomass to either physicochemical or biological deconstruction has been hampering its broad industrial utilization and, consequently, the transition to a sustainable bioeconomy. The gut system from herbivores are formidable bioreactors in nature for lignocellulose breakdown and the diverse ecological niches where herbivores are found have led to the rise of a myriad of molecular strategies to cope with the sheer complexity of plant polysaccharides. This study illuminates how the unexplored microbiota of the largest living rodent, capybara, found in Pantanal wetlands and the Amazon basin, can efficiently depolymerize and utilize lignocellulosic biomass.

Results

Here, we have elucidated the gut microbial structure and composition of the semiaquatic herbivorous capybara through multi-omics approaches. Metabolic reconstruction of this microbiota showed that cellulose degradation is chiefly performed by Fibrobacter bacteria, whereas hemicelluloses and pectins are processed by a broad arsenal of Carbohydrate-Active enZymes (CAZymes) organized in polysaccharide utilization loci (PULs) identified in multiple metagenome-assembled genomes from the phylum Bacteroidetes. Furthermore, metabolomics analysis showed short chain fatty acids as major fermentation products, which are key markers of digestion performance of plant polysaccharides. Exploring the genomic dark matter of this gut microbial community, two novel CAZymes families were unveiled including a glycoside hydrolase family of β-galactosidases and a carbohydrate-binding module family involved in xylan binding that establishes an unprecedented three-dimensional fold among associated modules to CAZymes.

Conclusions

Our results unveil how the capybara gut microbiota orchestrates the depolymerization and utilization of dietary plant polysaccharides, representing an untapped reservoir of new and intricate enzymatic strategies to overcome the recalcitrance of plant polysaccharides, a central challenge toward a circular and sustainable economy.

Introduction

The diverse symbiotic microbiota present in the digestive tract of herbivores has been an overwhelming source of intricate enzymatic mechanisms for lignocellulose deconstruction (1–4). In particular, the microbiota of foregut (rumen) fermenters has served for decades as a model system (5, 6), which led to the discovery of sophisticated and intriguing systems to degrade complex plant fibers such as the multi-enzyme complexes cellulosomes from *Ruminococcus flavefaciens* (7) and the efficient and distinguishing cellulose degradation system of *Fibrobacter succinogenes* (8).

Hindgut fermenters represent another class of herbivores probably as rich as ruminants in enzymatic mechanisms for the breakdown of recalcitrant plant glycans, since they are able to efficiently utilize low-quality forage (9). Similar to foregut fermenters, the digestion is accomplished by a symbiotic microbial community, but it occurs in a single and enlarged fermentation chamber (10). These monogastric herbivores comprise a vast range of animals from massive mammals such as elephants, rhinos and horses to small animals exemplified by rabbits and semiaquatic rodents (11). In addition, they are spread over a myriad of ecological niches in all continents such as rain forests, savannas, grasslands, swamps, highlands and deserts, suggesting to have evolved highly specialized molecular strategies to overcome the sheer complexity and diversity of plant glycans in these environments.

Capybara (*Hydrochoerus hydrochaeris*) is the largest living rodent throughout found in Pantanal wetlands and the Amazon basin, which is also known as “Master of the grasses” due to its diet. In this animal, the fermentation takes place in the cecum, which corresponds to almost three quarters of the digestive tract, reaching a digestive efficiency comparable to that of ruminants (12). Preliminary characterization of capybara cecum microbiome indicated that the most abundant commensal microbes were from Firmicutes and Proteobacteria phyla, with an unusual low abundance of Bacteroidetes (13), not expected for a typical mammalian hindgut fermenter. This initial study led us to the hypothesis that this gut microbiome may employ diversified metabolic capabilities and enzymes for depolymerization and fermentation of complex dietary fibers to provide host nutrition. Moreover, capybara animals dwelling the Southeast region in Brazil for decades have incorporated the industrially relevant sugarcane in their diet (14), which makes their cecal microbiome an especially attractive system for lignocellulose depolymerization.

To address this knowledge gap, we performed an integrated multi-omics approach to reveal the community structure and composition, Carbohydrate Active EnZymes (CAZymes) repertoire and metabolic pathways of the gut microbiota from native capybara animals dwelling the Brazilian Southeast region. Furthermore, by combining carbohydrate enzymology, X-ray crystallography and mutagenesis, two new CAZy families involved in plant glycan deconstruction were discovered, highlighting the potential of the capybara´s gut microbiome as a reservoir of unprecedented enzymatic systems for carbohydrate processing.

Results

Taxonomic structure and composition of the capybara gut microbiota

To explore the microbial community structure, membership, and metabolic exchange of capybara gut microbiome, we collected replicated fresh samples from the cecum and rectum from three wild female animals. Herein we combined several culture-independent omics approaches including 16S rRNA targeted sequencing to access the community structure; whole shotgun metagenome sequencing (MG) to reveal the community genetic and functional profile; metatranscriptomic RNA sequencing (MT) to determine the

expression level of genes; and NMR-based metabolomics to elucidate the small molecules profile of this specialized community adapted to degrade recalcitrant plant polysaccharides.

The capybara gut microbiome taxonomic structure is dominated by Bacteria, with only a small fraction of reads corresponding to Archaea (16S: 1.07%, MG: 0.40% and MT: 1.55%), and Fungi (MG 0.12% and MT: 0.32%). The most abundant bacteria found in capybara gut microbiota based on 16S taxonomy analysis were members of the Firmicutes (mean ± sd: 35.8 ± 12.4%), Bacteroidetes (31.5 ± 9.8%), followed by Fusobacteria (15.3 ± 5.4%) and Proteobacteria (8.4 ± 5.4%) (Fig. 1A). A similar taxonomic distribution was observed for 16S rRNA reads recovered from metagenome (16S_MG), MG and MT datasets (correlation coefficient r = 0.96 for MG:MT, r = 0.74 for 16S_MG:MG and r = 0.71 for 16S_MG:MT, $P < 0.05$). Bacteroidetes, Fusobacteria and Proteobacteria phyla were significantly more abundant in MG than in MT considering cecal samples (Fig. 1B), whereas Euryarchaeota (MT/MG ratio expressed as mean ± sd: 2.3 ± 0.6), Fibrobacteres (1.7 ± 0.9) and Spirochaetes (1.6 +/- 0.4) were more represented in MT than in MG data in cecal samples (Fig. 1B). The higher MT/MG ratios for these phyla indicate that these microorganisms were more active in the time of samples collection, and thus that they may be important players in this gut compartment.

Binning of MG assembled contigs based on tetranucleotide frequency and coverage profile resulted in the building of 79 unique Metagenome Assembled Genomes (MAGs), with completeness > 55% and contamination < 15% (Supplementary Table 1); among those, 24 were considered of high quality (completeness > 90% and contamination < 5%) and 50 medium-quality (completeness > 50% and contamination < 10%), according to parameters suggested by Bowers *et al*. 2017 (15). Taxonomy classification indicates that 35 of the recovered MAGs belong to the Firmicutes phylum, including six from the *Erysipelotrichaceae* family and eight from the *Lachnospirales* family. The second most abundant group was the Bacteroidetes, with 30 MAGs classified in the Bacteroidales order (Supplementary Table 1). Although only two genomes from Fusobacteria and Proteobacteria were recovered from MG, the most abundant OTUs identified by 16S analysis were classified as Fusobacteria and Proteobacteria with a relative abundance of 20% and 18%, respectively, (Figure S1), pointing to a central role of those species in this environment.

A dominance of Firmicutes, Proteobacteria, Bacteroidetes and Tenericutes was observed in the hindgut microbiomes of other herbivores such as *Castor fiber, Castor canadensis*, horse, rabbit and koala (16–20). Further, microbiota analysis of domesticated herbivores including hindgut fermenters, ruminants and monogastric animals revealed Firmicutes as the dominant phylum (53.11, 63.35 and 52.27% respectively), followed by Bacteroidetes (31.36, 20.95 and 26.95%, respectively). Although the dominance of Bacteroidetes and Firmicutes is a general feature of mammalian gut microbiomes, the microbiota of native Brazilian capybara differs from other hindgut fermenters and ruminants, mainly due to a reduced abundance of Firmicutes (35%) along with a higher abundance of Fusobacteria (15%) and Proteobacteria (8%) (21). The increased presence of Fusobacteria can be associated with the production of butyrate, a short-chain fatty acid that is often the end-product of carbohydrate fermentation (22). On the other hand, and in spite the high polysaccharide diet, the lower abundance of the Firmicutes in the capybara

microbiome may point to strategies for lignocellulose utilization distinct from those typically found in other hindgut herbivores and ruminants.

Metabolic profiling indicates high performance on the conversion of dietary fibers

Recalcitrant glycans found in diet components such as cellulose, hemicellulose and pectins are processed via anaerobic microbial fermentation to produce a wide range of metabolites, reflecting the diversity of substrates available in the digestive tract of herbivores, as well as the biochemical potential of the gut microbiota. The major fermentation products detected in the capybara gut by NMR spectroscopy-based metabolomics, were short-chain fatty acids (SCFAs) such as acetate, propionate, and butyrate, among more than 40 metabolites measured (Supplementary Table 2). SCFAs were detected in high concentration in both cecal and rectal samples. The most abundant metabolites were acetate (mean ± SD: 74.83 ± 22.17 and 30.40 ± 22.76 μM), propionate (31.0 ± 6.67 and 15.98 ± 12.8 μM) and butyrate (23.30 ± 5.63 and 8.35 ± 12.83 μM) in cecal and rectal samples, respectively (Supplementary Table 2). These SCFA ratios indicate a forage-based diet and are similar to that observed for ruminants (23, 24).

The MG and MT datasets were analyzed to describe the microorganisms and metabolic pathways associated to fermentation and SCFA production (Supplementary Fig. 2A). Genes related to pyruvate fermentation were highly abundant in both MG and MT data for cecal and rectal samples and the microbiota related to this pathway was dominated by Firmicutes, Bacteroidetes and Fusobacteria (Supplementary Fig. 2B). Metabolic pathways reconstruction of the 79 unique genomes recovered from capybara gut microbiome was conducted to further investigate the contribution of individual microorganisms to SCFAs production (Fig. 2). This analysis indicates that acetate can theoretically be produced by any of the bacterial genomes recovered from capybara gut microbiome, in agreement with the high abundance of this metabolite in both cecal and rectal samples (Table S2). Butyrate is known to be produced mainly by Firmicutes and the analysis of the key genes involved in the final steps of this pathway including butyryl-CoA:acetate CoA-transferase *atoA/D* genes, *butK* and *ptb* genes encoding butyrate kinase (EC 2.7.2.7) and phosphotransbutyrylase (EC 2.3.1.19), respectively, showed that Firmicutes *Ileibacterium* sp. MAG6 and *Megasphaera* sp. MAG33 are likely the major butyrate-producing bacteria in the capybara gut since they present the highest expression of *atoA/D* genes (Figure S3 and Table S3). Other bacteria, for instance the Bacteroidetes *Marinilabiliaceae* MAG47 and Fusobacteria MAG38 and MAG39 also presented co-localized genes *atoA*/ *atoD* and *ptb*/ *butK*, suggesting that they may also contribute to butyrate production (Figure S3 and Table S3).

In order to verify the distribution of the pathways for propionate production within the capybara gut microbiota, key genes from each pathway (acrylate, propanediol or succinate) were analyzed (25). Lactoyl-CoA and propane-1,2-diol, intermediates from acrylate and propanediol pathways respectively, were not identified in the metabolic reconstruction of any of the genomes recovered from capybara gut (Fig. 2). On the other hand, the succinate pathway, assessed by the *mmdA* gene encoding methylmalonyl coA decarboxylase, was widespread mainly among Bacteroidetes, but also detected in some Firmicutes

and Fusobacteria genomes (Figure S3 and Table S3), indicating that the main substrate used by capybara gut microorganisms to propionate production are probably hexoses and pentoses. Furthermore, the proportion of propionate detected in the gut capybara gut correlates (R = 0.77 and p = 0.07) with the relative abundance of Bacteroidetes, reinforcing that succinate pathway of this phylum is the major source of propionate production in capybara gut.

A few gut microorganisms are known to produce both propionate and butyrate, such as *Roseburia inulinivorans, Coprococcus catus* and *Eubacterium hallii* (26, 27). Other microorganisms able to produce acetate, butyrate and propionate as metabolic end products are *Megasphaera* sp. NM10, BL7 and *M. elsdenii* (28). According to metabolic reconstruction analysis, butyrate and propionate were predicted to be present concomitantly in 15 genomes (Fig. 2) and *Megasphaera* sp. MAG33 shares ci. 95% identity to the ruminal *M. elsdenii* suggesting similar metabolic capabilities. These observations reinforce the idea that the capybara microbiome is a promising source of novel species with diversified metabolic functions, with great potential for the breakdown of dietary structural carbohydrates as the high SCFA production are common markers of digestion performance of recalcitrant plant fibers (29).

Capybara gut microbiome strategies for the breakdown of dietary polysaccharides

The capacity of capybara to convert lignocellulosic materials into SCFAs is determined by the genomic potential associated with Carbohydrate-Active enZymes (CAZymes) of the gut microbiota. A total of 6,132 putative CAZymes encoding genes from 105 Glycoside Hydrolases (GH) and 10 Polysaccharide Lyases (PL) families were identified, of which 456 genes presented a modular architecture (Figure S4 and Table S4). The most abundant CAZymes identified are plant cell wall-degrading enzymes from families GH3, GH2 and GH1 (by decreasing abundance) that encompass diversified activities including β-glucosidases, β-xylosidases, β-galactosidases and β-mannosidases, among others. These enzymes are often associated with the later steps in the degradation cascade of several plant polysaccharides such as cellulose, heteroxylans, mixed-linkage β-glucans and β-mannans. Moreover, it has already been reported that these families are highly abundant in several host-associated gut microbiomes such as human, mouse, swine, and cattle rumen (30), probably due to their broad functions.

As sugarcane is part of the capybara diet dwelling Brazil Southeast region, it was expected that its microbiota would be able to use the easily metabolizable sugar sucrose. In capybara gut CAZymes arsenal, invertases from GH32 family were identified in a proportion of approx. 1.5%, which is similar to that reported for several gut microbiomes from ruminants to humans (30, 31). It is worth to mention that in the sequenced genome of capybara itself there is no gene encoding GH32 invertases, which holds for all mammals sequenced to date. Further analysis of MG and MT datasets, revealed a high abundance of GH32 enzymes in MG only (Figure S4), which led us to the hypothesis that, although the gut microbiome has the genomic potential to metabolize sucrose, the capybara was digesting more recalcitrant components of its diet at the time of sample collection.

One of the main dietary polysaccharides of capybara is cellulose, which is highly resistant to microbial degradation due to its chemical and structural organization along with numerous intermolecular

interactions with a complex matrix of hemicelluloses, pectins and lignin. Neither cellulases from families GH6, GH7 and GH48, nor cellulosomes, assessed by the presence of cohesin and dockerin domains associated with cellulases, could be identified in capybara gut MG or MT datasets. This suggests that cellulose degradation in the capybara gut may be accomplished by endo-β-1,4-glucanases (EC 3.2.1.4) from families GH5 (subfamilies GH5_2, GH5_4, GH5_25 and GH5_37), GH8, GH9 and GH45, which were detected either as single domains or in multi-modular protein architectures. Interestingly, the most expressed genes putatively encoding endo-β-1,4-glucanases detected in capybara gut microbiome belong to families GH5_2, GH8, GH9 and GH45 and were recovered from *Fibrobacter* genomes (Figure S5 and Table S5), indicating that these bacteria may be the major contributors to cellulose degradation in the capybara gut. *Fibrobacter succinogenes* is known as a highly efficient cellulolytic bacterium in the cow rumen (32). It is proposed that *F. succinogenes* utilizes a multi-protein complex to attach to cellulose fibers and secretes cellulases by the T9SS-dependent secretion system to enable cellulose breakdown into cellodextrins, which then would be imported into the periplasm for further degradation and utilization (33). The three *Fibrobacter* genomes recovered from capybara gut microbiome encode cellulases with a T9SS signal sequence as well as proteins for cellulose adhesion including tetratricopeptide, fibro-slime, OmpA and pilin proteins, as reported for *F. succinogenes* (33). Furthermore, from the set of 347 proteins observed in the outer membrane vesicles (OMVs) from *F. succinogenes* (34), we have identified 262 with sequence identity ranging from 30–99%. These observations suggest that typical *Fibrobacter* mechanisms, fundamentally relying on cell surface adhesion and OMVs, are central for cellulose degradation in the capybara gut.

Hemicelluloses and pectins are also important polysaccharides in the diet of capybaras and 30 Bacteroidetes genomes were recovered from capybara gut microbiota. Bacteroidetes are known to possess highly diversified carbohydrate degradation capabilities, many of them encoded as polysaccharide utilization loci (PULs), which are clusters of genes encoding CAZymes, SusCD-like transporter and regulators. Around 120 predicted PULs and 150 Clusters of CAZymes (CCs) were identified in our Bacteroidetes MAGs (Extended Data Fig. 1 and Table S6), and were compared to literature-derived PULs available in the PULDB database (35). PULs probably involved in the degradation of xylans and arabinoxylans – polysaccharides highly abundant in grasses including sugarcane – were identified in the genomes of *B. heparinolyticus* MAG 61 and Bacteroidota bacterium MAG40 (Fig. 3A), resembling PULs from *B. ovatus* (36). The strategies for the breakdown of mixed-linkage β-glucans are highly conserved in capybara and human microbiomes, with an exact same PUL organization encompassing GH16 and GH3 enzymes (Fig. 3A) (37). PULs involved in xyloglucan (XyG) degradation, a more recalcitrant hemicellulose, were identified in the Bacteroidaceae bacterium MAG53, featuring core hydrolases from families GH5_4, GH31 and GH9 (Fig. 3A). In *B. ovatus*, the XyG-PUL encodes other enzymes from GH43, GH3 and GH2 families (38), which were also detected in MAG53, albeit in distinct genomic regions. These enzymes may function as escorts for a complete depolymerization of XyGs similar to that reported for the saprophyte *Cellvibrio japonicus* (39). PULs predicted to act on mannose-containing glycans were also identified in the capybara gut microbiome (Fig. 3A), conserving the core genes GH26 (endo-β,1-4-mannanases) and GH130 (β-1,4-mannosylglucose phosphorylases) as described

for the human gut bacteria *B. fragilis* (40). Furthermore, a set of different PULs putatively enabling the degradation of other polysaccharides such as starch and pectins, were identified mainly present in Bacteroidaceae genomes (Fig. 3A and Table S6). For instance, PUL54 from Bacteroidaceae bacterium MAG51 involved in the degradation of homogalacturonan, a key component of sugarcane cell wall pectin (41), comprising enzymes from families GH105, GH43_10 and GH28 (Fig. 3A and Table S6) resembles the corresponding PUL from *B. ovatus* (36). However, a clear target substrate could not be defined for a large fraction of PULs predicted from Capybara gut microbiome (Table S6), in part due to intrinsic limitations of genome reconstruction from metagenomes, but also reflecting the variability, heterogeneity and insufficient knowledge of the structure and composition of the glycans present in the diet of wild capybaras. Nevertheless, our analyses highlight the importance of the Bacteroidetes phylum in the Capybara gut providing a diverse arsenal of enzymatic systems for the degradation and utilization of the main components of dietary carbohydrates.

Taken together, our results demonstrate that the capybara gut microbiota preferentially exploits a combination of free enzymes (rather than cellulosomes) containing a catalytic module either isolated or appended to CBMs or other catalytic modules to deconstruct dietary polysaccharides with a biochemical diversity provided by Bacteriodetes PULs/CCs and with *Fibrobacter* genera as workhorses for cellulose breakdown.

A new partner for an old acquaintance in heteroxylan degradation

Among the genomes recovered from capybara gut microbiome, *Prevotella* sp. MAG57 is the one with the largest number of CAZyme-encoding genes (Fig. 3B and Table S6). Phylogenetic analysis and whole genome comparison indicated that MAG57 is closely related to other uncultured genomes from the *Prevotella* genus recovered from capybara and from the UBA project (42) from sheep, elephant and mice gut (Fig. 6A). Regarding sequence-based genomic comparisons, MAG57 has an average nucleotide identity (ANI) of 75% but with an alignment fraction < 60% to genomes selected across Bacteroidetes phylum, and thereby it most likely corresponds to a novel species (Figure S6B). Many different PULs and CAZyme clusters organizations were identified in MAG57, probably involved in the degradation and utilization of hemicelluloses and pectins (Table S6). In particular, a gene cluster with predicted GH10, GH43 and GH97 members drew our attention as putatively acting on arabinoxylans, an abundant hemicellulose in secondary cell walls of sugarcane and other grasses. In particular, its GH10 member appear to contain an unknown N-terminal domain extension with a predicted mass of approx. 45 kDa (Fig. 4A). Sequence analysis showed that this unusual N-terminal domain is also present in Bacteroidetes species derived from human, mouse, and elephant gut-associated species (Table S7). However, it displays no similarity to domains typically associated with GH10 members such as xylan-binding CBM22 and xylanase-specific CBM9.

To evaluate the function of this unconventional GH10 member (CapGH10), the full-length protein and its domains along with other GH members of the CC102 cluster were recombinantly expressed and characterized. The GH97 member (CapGH97) is a calcium-activated α-galactosidase, whereas the GH43

member is a highly active α-L-arabinofuranosidase (Figure. S7-S8 and Table 1), two critical activities to remove decorations of heteroxylans. The later belongs to subfamily GH43_12 and showed low sequence identity to other structurally characterized GH43 members [~ 34% with *Bacteroides ovatus* GH43a, PDB 5JOW (43)]. Structural elucidation by SeMet phasing (Table S8) revealed a two-domain architecture with a β-sandwich accessory domain tightly bound to the catalytic domain (Figure S8D). Distinct to all other GH43_12 members structurally characterized so far, in which the β-sandwich domain is composed only by C-terminal β-strands, the GH43_12 structure herein elucidated shows an N-terminal β-strand that integrates with C-terminal β-strands to form the β-sandwich domain (43–45) (Figure S8 D). It indicates a further level of structural complexity within the GH43 family that might be carefully considered when designing constructs and chimeras involving these instrumental enzymes for plant polysaccharides depolymerization. Structural comparisons with other GH43_12 arabinofuranosidases showed a highly conserved active-site pocket including all residues comprising − 1 subsite, which is in agreement with the specificity and action mode of CapGH43_12 (Figure S8 E-F).

Table 1
Kinetic parameters of CAZymes heterologously expressed in *E. coli* BL21.

Protein ID	CAZy family	Substrate	pH	T (°C)	K_M	k_{cat} (s^{-1})	k_{cat}/K_M
09512	GH97	pNP-α-D-Gal	7.0	35	8.43 ± 0.57 (mM)	34.1 ± 0.98	4.05
09513 (full-length)	GH10	Rye arabinoxylan	5.5	50	2.14 ± 0.44 (mg/mL)	127.7 ± 16.3	59.67
09513 (GH10 domain)	GH10	Rye arabinoxylan	5.5	55	1.93 ± 0.08 (mg/mL)	180.1 ± 5.3	93.31
		Xylan	5.5	55	1.69 ± 0.08 (mg/mL)	160.6 ± 5.22	95.03
09514	GH43_12	α-L-arabinofuranoside	6.5	35	2.74 ± 0.29 (mM)	151.19 ± 6.21	55.18
44807	GHXXX	pNP-β-D-Gal	7.5	45	0.57 ± 0.05 (mM)	17.6 ± 0.39	30.88
CBK67650.1 β-Gal Domain	GHXXX	pNP-β-D-Gal	7.5	45	1.19 ± 0.35 (mM)	29.85 ± 1.95	25.08

The GH10 domain of the CapGH10 protein was shown to be an endo-β-1,4-xylanase active on beechwood xylan and several arabinoxylans including high viscosity rye flour arabinoxylan (33 cSt), low viscosity wheat flour arabinoxylan (8 cSt), acid debranched wheat arabinoxylan (26% Ara and 22% Ara) and enzyme debranched wheat arabinoxylan (30% Ara). Kinetic analyses indicate that decorations present in rye arabinoxylan (arabinose/xylose ratio = 40/60) are not detrimental to the enzyme catalytic performance, exhibiting similar K_m and k_{cat} constants compared to xylan (Table 1 and Figure S9). The Xyn10Z enzyme from *Hungateiclostridium themocellum* ATCC 27405, sharing 36% of sequence identity

with CapGH10, is the closest characterized member so far, with high activity on xylan (46). The N-terminal region of Xyn10Z comprises a feruloyl esterase followed by a CBM6 domain, both of which are not present in CapGH10 (47). The CapGH10 N-terminus, comprising approximately 500 residues, showed only sequence similarity with uncharacterized proteins, with the closest homologs mostly presenting a GH10 domain with sequence identity around 37–44%, and further hypothetical proteins without the GH10 module but with a T9SS signal domain sharing ca. 30% sequence identity. Homologs with similar domain architecture, attached to the N-terminus of a GH10 module, were found in PULs from ruminal *Prevotella* sp. such as *Prevotella* sp. BP1-148, *Prevotella* sp. BP1-145, Prevotellaceae bacterium HUN156 and Prevotellaceae bacterium MN60. These PULs further comprise members from families GH97, GH43_29 + CBM6 and CE1 + CE6 + CBM48, and are likely targeting xylan-related polysaccharides.

The potential enzymatic activity of the isolated N-terminal domain of CapGH10 was assessed for over 30 different substrates including synthetic substrates, oligosaccharides, and polysaccharides (Supplementary Table 9), but no (hydrolase, lyase or esterase) activity was observed. Typical activities involved in heteroxylans breakdown including endo-β-1,4-xylanase, β-xylosidase, α-L-arabinofuranosidase, α-D-galactosidase, α-D-glucuronidase, 4-O-methyl-glucuronoyl methylesterase, feruloyl esterase and acetyl xylan esterase were assayed by distinct methods without the detection of product formation or substrate consumption. Under this perspective, we further interrogated the capacity of this N-terminal domain to bind potential substrates of its GH10 partner such as beechwood xylan and arabinoxylans using affinity gel electrophoresis (AGE). As shown in Fig. 4C, this domain can indeed interact with the substrates of the GH10 domain, suggesting that this N-terminal domain may target the CapGH10 catalytic domain to xylan polysaccharides (Fig. 4C).

To get further insights into the potential role of this unconventional N-terminal domain, its crystallographic structure was solved by SeMet phasing at 1.8 Å resolution (Table S8). The domain exhibits a parallel right-handed β-helix fold, consisting of 14 complete helical turns with two main short helices protruding from the β-helix backbone (Fig. 4B). The 14 helical turns are twisted and curved with a calcium ion between the 11th and 12th turns in an octahedral coordination sphere (Fig. 4B). This β-helix fold is observed in the clan GH-N of the GH superfamily, in the carbohydrate esterase CE8 and in several polysaccharide lyase (PL) families; however, structural comparisons with these CAZy families led to high rmsd values (> 3 Å), indicating poor three-dimensional conservation (Table S10). Despite that, structural superpositions were performed with CAZy families (GH28, GH91, PL6 and CE8) as an attempt to identify similarities of CapGH10 β-helix domain with the active sites of these enzymes. Neither the catalytically relevant residues nor the active site topology of these families are conserved in the CapGH10 β-helix domain (Extended Data Fig. 2). Besides the lack of all key catalytic residues, a long loop (G126-K140) in the CapGH10 β-helix domain also partially occludes the region corresponding to the active site in the GH28 enzymes (PDB ID 3JUR (48)) (Extended Data Fig. 2A). In comparison to family GH91 (PDB ID 2INU (49)), the two loops critical for catalytic activity, T2 and T3, are absent in CapGH10 β-helix domain (Extended Data Fig. 2B) and in the PL6 family (PL6, PDB ID 6QPS (50)), the Ca^{2+}-binding site essential for catalytic activity is not present in CapGH10 β-helix domain (Extended Data Fig. 2C). Despite there is a

cleft-like region in the CapGH10 β-helix domain near to the corresponding active site of the CE8 family (PDB ID 3UW0 (51), Extended Data Fig. 2D), the catalytic residues are not conserved and most residues populating this region in the CapGH10 β-helix domain are not even conserved within homologues, weakening the possibility of this region to be a catalytic center. Moreover, SAXS data (Figure S10) indicated that the CapGH10 β-helix domain is monomeric in solution, unlike the GH28 and GH91 families that rely on oligomerization to be functional. These structural analyses, and the lack of conservation of residues corresponding to the cleft-like region in CapGH10 β-helix domain homologues support the biochemical data that this domain is not catalytically active.

Considering aromatic and acidic residues as important platforms for carbohydrate interaction, mapping of the molecular surface of the CapGH10 β-helix domain led to two potential binding regions, one between turns 1–4 (region I) and another between turns 6–10 (region II). Therefore, residues Y62 and E82 from region I and residues E132, D133, Y193, E225, E247, Y279, E282, D360 and D365 from region II were mutated to alanine (Supplementary Fig. 11). Moreover, one mutation at the calcium-binding site (D344L) was evaluated to address whether calcium ion incorporation could be essential for carbohydrate binding. Mutations E247A and E282A severely impaired protein stability and led to the expression only in the insoluble fraction. Mutation D344L also affected protein stability in a less extent, but the arabinoxylan/xylan binding capacity was preserved (Figure S12). This result indicates that calcium ion has a structural relevance rather than a functional role in carbohydrate recognition. Among the other nine mutants, only Y62A and E82A, affected the migration pattern in AGE assays with beechwood xylan and rye arabinoxylan (Fig. 4C). Both residues are located at the region I, indicating that this region plays a role in carbohydrate binding. It is worth to mention that two aromatic residues located at the corresponding region of the GH28 active site, Y193 and Y279, did not alter the carbohydrate binding, being in agreement with no functional relevance of this region for CapGH10 β-helix domain. Combining the biochemical, structural and mutagenesis analyses, we would define CapGH10 β-helix domain as a CBM, therefore, establishing a novel structural scaffold in this superfamily and founding the new family CBMXX.

Taken together this unprecedented modular endo-β-1,4-xylanase along with the synergistic activities of other CC107 partners, we conclude that this cluster confers the ability to *Prevotella* sp. MAG57 to act on complex heteroxylans (Fig. 4D), a key function in the gut microbiome of capybara that have grasses as a major component in its diet.

A new GH family mined from the genomic dark matter of capybara microbiome

The combined MG and MT analysis of capybara gut microbiome revealed several expressed genes annotated as hypothetical proteins. Some of these genes presented extremely remote similarity to CAZy members, with percentage of sequence identity ranging from 10 to 20%, suggesting a potential function in the processing of plant polysaccharides, but requiring confirmation by functional investigation (Table S11). Aiming to uncover the activity of these proteins, synthesized ORFs were expressed and subjected to biochemical assays employing a diverse set of synthetic, poly- and oligosaccharides substrates.

One of these proteins (SEQ ID PBMDCECB_44807, named here CapGHXXX) was active on *p*-nitrophenyl-β-D-galactopyranoside (*p*NP-β-D-Gal), and its kinetic parameters were determined (Table 1 and Figure S13). CapGHXXX orthologues are present in Actinobacteria, Firmicutes, Verrucomicrobia and mainly in Bacteroidetes genomes recovered from diverse sources such as rumen, feces, gut and oral microbiota (Table S12), being the closest sequence from a rumen-derived genome (UBA2817) from the uncultured RC9 group (42). Sequence analysis showed that CapGHXXX is distantly related to families GH5 and GH30 (Fig. 5A) and protein threading indicates a TIM barrel fold (Supplementary Fig. 14), suggesting that this novel GH family belongs to the clan GH-A. To further explore this GH family, the enzyme CBK67650.1 (SEQ ID BXY_26070) from *B. xylanisolvens*, which shares 46% sequence identity with CapGHXXX, was synthesized, produced and biochemically characterized (Table 1). This second member also showed β-galactosidase activity that strengthens at biochemical level the establishment of this new GH family.

In the genome of Bacteroidota bacterium MAG42 recovered from Capybara gut, CapGHXXX is found in a putative PUL additionally comprising enzymes from families GH2 and GH78. A similar PUL organization was predicted in the genome of *Bacteroidetes* sp. 1_1_30 recovered from human gut, which yet harbors enzymes from GH36, CE7 and PL8_2 families. It is noteworthy that CapGHXXX is often found fused appended to a GH36 module or in PULs also having GH36 members such as in *B. xylanisolvens* and *Prevotella dentalis*, recovered from stool and oral cavity, respectively (Fig. 5B), indicating a synergistic relationship between these families. Moreover, these families are also commonly found along with GH78 α-L-rhamnosidases in the PUL context. In the genome of the Bacteroidales bacterium UBA2817, a GHXXX member is appended to a GH78 module carrying a CBM67, both targeting rhamnogalacturonans (Fig. 5B). These observations suggest that GHXXX could act on β-linked galactosyl residues in pectic polysaccharides. Further studies in the PUL context are required to shed light on their biological role in complex gut environments.

Discussion

The capybara (*Hydrochoerus hydrochaeris*), also known as “Master of the grasses”, is the largest rodent living on earth, dwelling Pantanal wetlands and the forests and plains of the Amazon basin. This semi-aquatic herbivore is a hindgut fermenter with an enlarged cecum that can efficiently degrade and utilize recalcitrant plant polysaccharides by microbial processes so far unexplored. Interestingly, in the Southeast region of Brazil these animals have incorporated sugarcane in their diet for decades, raising the possibility that their gut microbiome has been shaped by this biomass of great industrial relevance.

Multi-omics analysis of the capybara gut microbiome revealed that carbohydrate processing resides on an elaborated arsenal of CAZymes from a diversified set of microorganisms from Bacteroidetes, Firmicutes, Fibrobacteres and Fusobacteria phyla, which yet exhibit distinct metabolic pathways to convert dietary fibers into SCFAs, a major energy source for the host (Fig. 6). Our analyses indicate that *Fibrobacter* bacteria are probably the workhorses for cellulose breakdown, involving the orchestrated action of diverse single-domain and as well as modular endo-β-1,4-glucanases from families GH5, GH9 and GH45. On the other hand, the degradation of hemicelluloses and pectins is catalyzed by an intricate

and broad repertoire of PULs/CCs observed in the assembled Bacteroidetes genomes targeting complex heteroxylans, xyloglucans, mixed-linkage β-glucans, homogalacturonans and rhamnogalacturonans, which are abundant polysaccharides in grasses, in particular sugarcane (41). It is noteworthy that many putative PULs/CCs identified in these MAGs only showed similar organization with predicted PULs for which a defined target substrate is unknown, pointing to a number of yet unexplored strategies for glycan processing in this microbiome.

Metagenomics along with metatranscriptomics also revealed a notable number of genes remotely related to known CAZy families or modular architectures comprising unknown domains, leading us to further explore this genomic dark matter through carbohydrate and structural enzymology. A cluster of CAZymes specialized in complex heteroxylans from a novel *Prevotella* bacterium contains an unconventional modular GH10 endo-β-1,4-xylanase, featuring a novel CBM family targeting xylans and arabinoxylans. This CBM family exhibits an original fold among the 87 known CBM families and an unusual high molecular weight for a typical CBM, expanding the known three-dimensional architectures in this superfamily. Furthermore, a new GH family has been established with the discovery of two enzymes exhibiting β-galactosidase activity but insufficient sequence similarity for inclusion to previously described CAZy families. This new family is phylogenetically and structurally related to the large GH-A clan.

Together, these results provide an unprecedented and comprehensive understanding of the enzymatic apparatus in the capybara gut microbiome specialized in the breakdown of lignocellulosic biomass.

Conclusion

Multi-omics analysis has unveiled the biochemical and metabolic pathways employed by the gut microbiota from the Amazon monogastric semi-aquatic herbivore, capybara, for the breakdown and utilization of recalcitrant dietary polysaccharides. This microbial community combines the unique cellulolytic machinery featured by Fibrobacteres and the diverse and elaborated PULs found in Bacteroidetes to efficiently depolymerize lignocellulosic biomass. Structural and functional investigation of proteins and PULs identified in the genomic dark matter of this microbiota uncovered new CAZy families, highlighting its great potential as source of novel molecular systems for the processing of plant polysaccharides. These findings expand our current understanding about gut microbial strategies to overcome the recalcitrance of lignocellulosic biomass, which might be utilized in biorefineries for the valorization of agroindustrial residues.

Material And Methods

Ethics statement

This study was carried out in strict accordance with the Animal Management Rule of the Brazilian Ministry of Environment (Documentation Sisbio 59826-1). The samples were obtained from three

euthanized animals, as a measure of management of Rocky Mountain Spotted Fever (RMSF) hosts, collected in Tatuí/São Paulo State, Brazil, in September 2017. After euthanasia, 20 g of intestinal contents were collected from the cecal and rectal of each animal. All samples were placed in sterile containers and immediately frozen in liquid nitrogen. Samples were kept at −80°C until processing.

Microbial DNA extraction

Samples of cecal and rectal contents were frozen in liquid nitrogen and pulverized with an oscillating ball mill (MM400; Retsch Inc., Newtown, PA). The homogenized samples were used for microbial DNA extraction according to the protocol described for (*52*) with modifications. Briefly, 0.25 g of sample was transferred to Lysing Matrix E Tube – Kit FastDNA Spin Kit for Soil (MP Biomedical, Inc.). For cell lysis, 1 ml RBB+C buffer was added in each sample, followed by homogenization in a FastPrep® FP120 instrument (MP Biomedical, Inc.). The precipitation of nucleic acids was obtained with the addition of a solution of ammonium acetate (10 M). The samples were incubated on ice for 30 minutes, after that, centrifugation was performed at 4°C for 10 min at 16,000×g. The nucleic acids pellet was recovered and washed with 70% ethanol, followed by drying at room temperature. The nucleic acid pellet was dissolved in 75 µL of autoclaved ultrapure water. RNA was removal in each sample with the addition of DNase-free RNase (10 mg/mL). DNA purification was performed using PowerClean® DNA Clean-Up Kit (Mo Bio Laboratories, USA) according to the manufacturer's protocol. Finally, electrophoresis using 0.8 % agarose gel was used to separate the DNA fragments and check the DNA quality. The DNA solution was stored at -20°C.

RNA extraction and mRNA enrichment

The samples homogenized with an oscillating ball mill were also used in this step. Briefly, 500 mg of sample was used for total RNA extraction by a Trizol protocol and FastRNA® Pro Green Kit (MP Biomedicals), according to the manufacturer's instructions. The total RNA samples were treated with a blended the Ribo-Zero™ Magnetic Kit*- (Bacteria) and Ribo-Zero™ Magnetic Kit* - (Human/mouse/rat) (Epicenter, Madison, WI, USA) to remove ribosomal RNA (rRNA) from total RNA and enrichment of mRNA. Approximately, 2.200 ng of total RNA was mixed with the blended Ribo-zero rRNA removal solution and incubated at 68°C for 10 min. After that, the reactions of RNA/rRNA were incubated 5 min at 50 °C with magnetic beads to remove the hybridized rRNA molecules from the mRNA. Following, the tubes were placed over the magnetic tube rack for 5 min to separate the beads from solution. The supernatant content the rRNA-depleted was carefully transferred to another RNase-free tube. The purification of rRNA-depleted was performed adding 200 µl of freshly prepared 80% ethanol to the tube while in the magnetic rack. The solution was incubated at room temperature for 5 minutes, and then the ethanol was discarded. The procedures were repeated twice, after that the samples were dried for 15 minutes at room temperature. The tubes were removed from the magnetic rack, followed for the addition of 11 µl of RNase-Free water and incubation for 2 minutes. Finally, the tubes were placed again over the magnetic tube rack, and the supernatant was collected. The total RNA obtained by magnetic beads procedure described above were used for sequencing.

Microbial community structure and diversity analysis

Capybara gut microbial community structure and diversity was investigated via high-throughput sequencing of 16S rRNA V4 region (LNBR, Brazil). The amplification of the 16S rRNA V4 region was done using the 515F (5'-GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) primers (*53*). Sequencing was performed on an MiSeq Sequencing System (Illumina Inc., USA) with the V3 kit, 600 Cycles, in paired-end sequencing with 2x300bp. The ZymoBIOMICS™ Microbial Community DNA Standard (D6305) from Zymo Research (Irvine, CA, USA), with eight phylogenetically distant bacterial strains (3 gram-negative and 5 gram-positive) and 2 yeasts, was included as a positive control to evaluate possible bias in libraries construction, sequencing and bioinformatics analysis. For taxonomy analysis, paired-end reads were quality checked using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and filtered using Trimmomatic v.0.36 (*54*), to remove adapters and low-quality reads. Filtered paired-end reads were merged using fastq_mergepairs function of Usearch v.10 package (*55*). Prior to OTUs clustering, primer sequences were removed using fastx_truncate function, reads were dereplicated and singletons were discarded. OTUs were clustered using the UPARSE unoise3 function. Prokaryotic and Eukaryotic taxonomy assignment was performed using sintax function with RDP database v16 (*56*) for 16S. Further analyses were performed using phyloseq v1.20 R package on R Studio.

Metagenome and metatranscriptome sequencing

Total DNA and RNA were extracted as described previously (item 2.2 and 2.3 respectively). Metagenomic libraries were prepared using the Nextera Library Preparation kit (Illumina), while metatranscriptomics libraries were prepared using the TruSeq Stranded total RNA Library Prep Kit (Illumina). Libraries concentration were measured through quantitative qPCR using KAPA Library Quantification Kit Illumina Platforms and assayed for quality using BioAnalyzer (Agilent Technologies, Palo Alto, Calif.). MG and MT libraries were paired end sequenced in two runs (2 × 100 bp) on the Illumina HiSeq 2500 platform at NGS sequencing facility available at LNBR (Brazil). The cecal and rectal gDNA were homogenized into a single sample and sequenced on a MinION sequencing device from Oxford Nanopore Technologies Inc. (ONT), to generate long-reads. About 1 µg of ultra-long high molecular weight gDNA from the homogenized samples was require for library preparation. The library was prepared with the SQK-LSK109 Kit (ONT) followed the standard protocols from Oxford Nanopore Technologies. The MinION sequencer run was performed on a Flow cell R9 version for ~22 hours at NGS sequencing facility available at LNBR (Brazil), generating around 3Gb of long reads.

Metagenome and metatranscriptome analysis

Metagenome and Metatranscriptome raw sequences were quality checked and trimmed as described above, MT reads were also analyzed using SortmeRNA to remove rRNA reads, and then both MG and MT reads were taxonomically classified using Kaiju (*57*). For functional analysis, the MG trimmed reads were *de novo* assembled using IDBA_UD (version 1.1.1) with the pre-correction parameter and k-mer size from 20 to 60 (*58*). Furthermore, the assembled metagenome was binned using CONCOCT v.0.4.0 (*59*) and

MaxBin 2.0 (*60*) to recover putative genomes from the metagenomic data. The binned genomes were dereplicated to remove redundancies using dRep and analyzed using CheckM v1.0.6 (*61*) to determine the completeness and contamination ratios of these genomes. Long-reads sequencing were used for MAGs scaffolding using SSPACE-long-reads v1.1. Genomes with completeness smaller than 55% and more than 15% contamination rate were discarded. Gene prediction and annotation of both the recovered genomes and the co-assembly was performed using Prokka v.1.11 with the meta parameter (*62*). KEGG pathways and Kegg Orthologous (KOs) annotation were performed using KOFAM (*63*) and Functional Ontology Assignments for Metagenomes (FOAM) database(*64*); CAZymes annotation was performed according to CAZy database pipelines (*65*). Furthermore, MG and MT reads were mapped to the assembled metagenome and recovered MAGs using Kallisto v. 0.46.1 (*66*) to estimate the coverage/abundance of protein coding genes in cecal and rectal samples. Normalized abundance was estimated based on the count/number of reads per kilobase per million mapped reads expressed as TPM.

Phylogenetic analysis and metabolic reconstruction

MAGs whole genome phylogenetic analysis was performed using the pipeline of PhyloPhlAn (*67*). To further assign taxonomy to the recovered genomes GTDB-tk tool was used (*68*). Phylogenetic analysis of the MAG57 and reference Bacteroidetes type strains was performed using concatenated 92 single copy core genes according to UBCG method (*69*). CAZymes phylogenetic tree was conducted using the catalytic domain of each family, aligned with MAFFT (*70*), by maximum likelihood methods employing RAxML software (*71*) with 1,000 rapid bootstrap inferences and LG as the substitution model.

To perform the metabolic pathways reconstruction of each recovered MAG, their annotation obtained from KOFAM database were filtered to keep only the top 5 hits of each protein with e-value below the 1e-5 threshold. These filtered annotations were then supplied to the Annotation of Metabolite Origins (AMON) tool (*72*), which based on the KOs annotated in the MAGs predicts the putative metabolites that can be generated.

Metabolite Profiling

Approximately 30 mg of the dry cecum and rectum contents sample and 300 µL of solution 2:1 (methanol: chloroform) were successively added to a tube. The mixture was sonicated for 1 minutes (15 second on, 10 second rest) and placed at 4°C for 15 minutes. Next, 300 µL of solution 1:1 (methanol: ultrapure water) was added. Followed by centrifugation at 16,000 g and 4°C for 20 min. The supernatant was transferred to a new tube. The supernatant samples were dried in CentriVap Solvent System – Labconco (Labconco Corporation, Kansas City, MO). Samples were diluted to 630 µL by addition of D2O, 70 µL of sodium phosphate buffer (final concentration 0.1 M) containing dimethyl-silapentane-sulfonate (final concentration 0.5 mM) for NMR chemical shift reference and concentration calibration. The samples were filtrated in a syringe filter with a 0.22 µm pore size hydrophilic polyethersulfone (PES) membrane. The final volume of filtrate ranged from 500 to 650 µL. 1H NMR spectra of samples were acquired using a Varian Inova NMR spectrometer (Agilent Technologies Inc., Santa Clara, USA) equipped with a 5 mm triple resonance cold probe and operating at a 1H resonance frequency of 599.84 MHz and

constant temperature of 298 K (25 °C). A total of 1024 free induction decays were collected with 32-k data points over a spectral width of 16 ppm. A 1.5-s relaxation delay was incorporated between scans, during which a continual water presaturation radio frequency (RF) field was applied. Spectral phase and baseline corrections, as well as the identification and quantification of metabolites present in samples, were performed using Chenomx NMR Suite 7.6 software (Chenomx Inc., Edmonton, Canada).

Protein expression and purification

Protein expression and purification were conducted as reported in (*73*). Briefly, *E. coli* BL21 strain was transformed with target genes subcloned into pET28a in frame to a 6xHisTag at N-terminus. LB medium [0.5% (w/v) yeast extract, 1% (w/v) tryptone, 1% (v/v) sodium chloride] was employed for protein expression in presence of according antibiotics for transformant selection. Culture growths were conducted at 37°C until $O.D._{600nm}$ around 0.8 and then, protein expression was targeted by addition of 0.2 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) (Sigma Aldrich). Protein expression was conducted at 18ºC for 16 hours and pelleted cells were further employed for protein purification.

Protein purifications were performed by nickel affinity chromatography followed by size exclusion chromatography. Briefly, pelleted cells were resuspended in saline-phosphate buffer (20 mM sodium phosphate, 500 mM NaCl, pH 7.5) with addition of 5 mM imidazole, 1 mM phenylmethylsulfonyl fluoride (PMSF), 5 mM benzamidine and 0.1 mg ml^{-1} lysozyme. Cells were disrupted by sonication and soluble protein lysates were applied to a 5-ml HiTrap Chelating HP column (GE Healthcare). Target proteins were eluted in an imidazole gradient up to 0.5 M. 6xHisTag were cleaved out of target protein by 1% (w/w) trypsin (catalog no. T1426, Sigma Aldrich). Target proteins were further purified by size exclusion chromatography with a HiLoad 16/600 Superdex 75 pg column (GE Healthcare) equilibrated with 20 mM sodium phosphate, 150 mM NaCl, pH 7.5. Purified proteins were evaluated by dynamic light scattering (DLS) and samples with low polydispersity (<20%) were further employed in biochemical and biophysical experiments. Samples were analyzed by SDS-PAGE and binding capacity of the CBMXX wild type and mutants were evaluated by affinity gel electrophoresis according to (*74*).

Crystallization, diffraction data collection, and structure determination

Crystallization experiments were carried by the sitting-drop vapor-diffusion method, in both sitting and hanging drops at 18 °C. The best crystals of CBMXX grew in 20% ($w.v^{-1}$) PEG 6000, 0.1 M sodium acetate (pH 5,0) and 0.2 M sodium chloride. GH43_12 crystals were obtained in 14% ($w.v^{-1}$) PEG 10,000, 1% ($v.v^{-1}$) dioxane, 0.1 M ammonium acetate and 0.1 M BIS-TRIS (pH 5.5). For cryoprotection, crystals were equilibrated with the reservoir solution added of glycerol or PEG 400 (20% ($w.v^{-1}$)) prior to flash cooling. Diffraction datasets of GH43_12 were collected at the MX2 beamline from the Brazilian Synchrotron Light Laboratory (LNLS, Campinas, SP) and of CBMXX at the PROXIMA-2A beamline from the SOLEIL Synchrotron (SOLEIL, Gif-sur-Yvette Cedex, France). All datasets were indexed, integrated, merged, and scaled using XDS package (*75*). The CBMXX and GH43_12 structures were solved by single anomalous dispersion (SAD) method using the diffraction data from Se-Met crystals. The programs SHELXC,

SHELXD and SHELXE (*76*)) were employed for data preparation, anomalous scatters location and phase calculation, respectively. Initial models were built with the AutoBuild Wizard (*77*) from the Phenix package (*78*). All structures were refined with programs PHENIX.REFINE (*79*) and REFMAC (*80*), and the models were inspected and manually adjusted according to the computed σ_A-weighted ($2F_o$-F_c) and (F_o-F_c) electron density maps using COOT (*81*). TLS groups were calculated by TLSMD (*82*) applied to both refinements. All structures were evaluated by MolProbity (*83*) and PDBRedo server (*84*). Structure factors and atomic coordinates of CBMXX and GH43 enzyme were deposited in the Protein Data Bank (PDB) under the accession codes 7JVI and 7JVH, respectively. Data collection and refinement statistics are summarized in Supplementary Table 8. The GH43_12 structure was firstly solved in in the C2 space group with 2 molecules found in the asymmetric unit. However, we obtained unusual high values for Rwork/Rfree for the data resolution even after many refinement/validation cycles, which could indicate a wrong space group or a crystal pathology. The data was then re-processed in the P1 space group with 6 molecules in the asymmetric unit and small decrease in the R values was obtained. It was observed a poor density in some molecules (E and F) in comparison with the others, probably due to several orientations assumed; our analyses of the data indicate a possible partial rotational order-disorder pathology.

Small Angle X-ray Scattering (SAXS) - Data collection and analysis

SAXS data of CBMXX was collected at the SAXS1 beamline (Brazilian Synchrotron Light Laboratory, Campinas, Brazil) at protein concentration of 8.4 $mg.mL^{-1}$ in 20 mM Hepes buffer pH 7.5. Buffer scattering were recorded and subtracted from the corresponding protein scattering. SAXS patterns were integrated using Fit2D (*85*) and GNOM (*86*) was used to evaluate the pair-distance distribution functions p(r). Ab initio molecular envelopes were calculated from SAXS data with DAMMIN (*87*) and averaged models were generated from several runs using DAMAVER (*88*). Each final SAXS low-resolution model was superimposed to its respective protein crystal structure using the program SUPCOMB (*89*).

Biochemical assays

α-L-arabinofuranosidase and α-galactosidase activity was determined at 35 °C in 70 to 80 mmol $liter^{-1}$ M, pH 6.0 and 7.0 (McIlvaine buffer), respectively. Beta-galactosidase activity was determined at 45 °C in 70 to 80 mmol $liter^{-1}$ M, pH 7.5 (McIlvaine buffer). The reaction mixture containing *p*-nitrophenyl-α-L-arabinofuranoside (pNP-α-LAraF), *p*-nitrophenyl-α-D-galactopyranoside (pNP-α-D-Gal) or *p*-nitrophenyl-β-D-galactopyranoside (pNP-β-D-Gal) (Sigma-Aldrich) as substrate in a final volume of 0.1 ml. The enzymatic activity was initiated by addition of the enzyme, following of 60 minutes of the incubation at 35- or 45°C, after that the reactions were interrupted by adding 0.1 ml of satured sodium tetraborate solution. The released p-nitrophenol in each reaction was measured at 400 nm separately, using Infinite® 200 PRO microplate reader (TECAN Group Ltd., Männedorf, Switzerland). The hydrolysis of polymeric substrates (Supplementary Table 10) were evaluated using the 3,5-dinitrosalicylic acid method by determination of reducing sugar released (*90*).

Declarations

Ethics approval

This study was carried out in strict accordance with the Animal Management Rule of the Brazilian Ministry of Environment (Documentation Sisbio 59826-1).

Consent for publication

All authors approved the final version of the manuscript.

Data and materials availability

All data for this study can be found under the bioproject ID PRJNA563062. The 16S, metagenomic and metatranscriptome reads for cecal and rectal samples are available at SRA under the accession numbers SRR11852069-SRR11852086; SRR11852046-SRR11852057 and SRR11852097- SRR11852108, respectively (Supplementary Table 14). The MAGs can be found at GenBank under the accession numbers JABUSA000000000 - JABUVA000000000. Structural data have been deposited in the Protein Data Bank (https://www.rcsb. org/) under accession codes 7JVI (CapCBMXX) and 7JVH (CapGH43_12). All other data generated or analyzed during this study are included in this published article (and its Supplementary information files) or are available from the corresponding author on reasonable request.

Competing Interests

The authors declare no competing interests.

Funding

This research was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (grant no. 2015/26982-0 to M.T.M. and postdoctoral fellowship 2016/19995-0 to M.A.B.M) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant no. 306135/2016-7 to M.T.M., 408600/2018-7 to G.F.P, 439195/2016-0 to L.C, 150552/2017-3 and 142332/2017-8 to M.P.M.). L.C, “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES; Finance code 001)".

Authors' contributions

LC, GFP and MTM designed the study and wrote the paper.

GFP, LFM, NT, VL and BH performed the bioinformatics analyses.

LC, DAP, MPM and MC performed the 16S, metagenomics and metatranscriptomics experiments.

LC, MPM, MC, MND, RASP, WCG and CAS expressed and purified the enzymes and performed the functional characterization.

LC, MPM, MABM and WCG performed the structural analysis.

LC and MLS performed the metabolomics analysis.

All authors analyzed the results and approved the final version of the manuscript.

Acknowledgements

We acknowledge the LNLS for the provision of time on the MX2 and SAXS1 beamlines, the Brazilian Biosciences National Laboratory (LNBio) for the use of the crystallization (Robolab), NMR and spectroscopy facilities, and the Brazilian Biorenewables National Laboratory (LNBR) for the use of the characterization of macromolecules and next generation sequencing facilities. LNLS, LNBio and LNBR are operated by the Brazilian Center for Research in Energy and Materials for the Brazilian Ministry for Science, Technology, Innovations and Communications. We acknowledge SOLEIL for provision of synchrotron radiation facilities at PROXIMA-2A (proposal 20181915) and we would like to thank William Shepard and Martin Savko for assistance in using the beamline. We acknowledge the support of Alana H. S. Alvarenga in protein purification and solubilization assays.

References

1. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA. Polysaccharide utilization by gut bacteria: Potential for new insights from genomic analysis. Nat Rev Microbiol. 2008;6:121–31.
2. Morrison M, Pope PB, Denman SE, McSweeney CS. Plant biomass degradation by gut microbiomes: more of the same or something new? Curr Opin Biotechnol. 2009;20:358–63.
3. White BA, Lamed R, Bayer EA, Flint HJ. Biomass Utilization by Gut Microbiomes. Annu Rev Microbiol. 2014;68:279–96.
4. Kartzinel TR, Hsing JC, Musili PM, Brown BRP, Pringle RM, Covariation of diet and gut microbiome in African megafauna. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23588–23593 (2019).
5. Krause DO, Denman SE, Mackie RI, Morrison M, Rae AL, Attwood GT, McSweeney CS. Opportunities to improve fiber degradation in the rumen: Microbiology, ecology, and genomics. FEMS Microbiol Rev. 2003;27:663–93.
6. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science (80-). 2011;331:463–7.
7. Rincon MT, Ding SY, McCrae SI, Martin JC, Aurilia V, Lamed R, Shoham Y, Bayer EA, Flint HJ. Novel organization and divergent dockerin specificities in the cellulosome system of Ruminococcus flavefaciens. J Bacteriol. 2003;185:703–13.
8. Burnet MC, Dohnalkova AC, Neumann AP, Lipton MS, Smith RD, Suen G, Callister SJ. Evaluating Models of Cellulose Degradation by Fibrobacter succinogenes S85. PLoS One. 2015;10:e0143809.

9. Demment MW, Van Soest PJ. A nutritional explanation for body-size patterns of ruminant and nonruminant herbivores. Am Nat. 1985;125:641–72.
10. Stevens CE, Hume ID. Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients. Physiol Rev. 1998;78:393–427.
11. Sakaguchi E. Digestive strategies of small hindgut fermenters. Anim Sci J. 2003;74:327–37.
12. Kiani A, Clauss M, Ortmann S, Vendl C, Congdon ER, Herrera EA, Kreuzer M, Schwarm A. Digestive physiology of captive capybara (Hydrochoerus hydrochaeris). Zoo Biol. 2019;38:167–79.
13. García-Amado MA, Godoy-Vitorino F, Piceno YM, Tom LM, Andersen GL, Herrera EA. M. G. Domínguez-Bello, Bacterial Diversity in the Cecum of the World's Largest Living Rodent (Hydrochoerus hydrochaeris). Microb Ecol. 2012;63:719–25.
14. Polo G, Mera Acosta C, Labruna MB, Ferreira F, Brockmann D. Hosts mobility and spatial spread of Rickettsia rickettsii. PLOS Comput Biol. 2018;14:e1006636.
15. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725–31.
16. Armstrong Z, Mewis K, Liu F, Morgan-Lang C, Scofield M, Durno E, Chen HM, Mehr K, Withers SG, Hallam SJ. Metagenomics reveals functional synergy and novel polysaccharide utilization loci in the Castor canadensis fecal microbiome. 12, 2757–2769 (2018).
17. Barker CJ, Gillett A, Polkinghorne A, Timms P. Investigation of the koala (Phascolarctos cinereus) hindgut microbiome via 16S pyrosequencing. Vet Microbiol. 2013;167:554–64.
18. Morrison PK, Newbold CJ, Jones E, Worgan HJ, Grove-White DH, Dugdale AH, Barfoot C, Harris PA. C. M. Argo, The Equine Gastrointestinal Microbiome: Impacts of Age and Obesity. Front Microbiol. 2018;9:3017.
19. Pratama R, Schneider D, Böer T, Daniel R. First Insights Into Bacterial Gastrointestinal Tract Communities of the Eurasian Beaver (Castor fiber). Front Microbiol. 2019;10:1646.
20. Velasco-Galilea M, Piles M, Viñas M, Rafel O, González-Rodríguez O, Guivernau M, Sánchez JP. Rabbit Microbiota Changes Throughout the Intestinal Tract. Front Microbiol. 2018;9:2144.
21. O' Donnell MM, Harris HMB, Ross RP. P. W. O'Toole, Core fecal microbiota of domesticated herbivorous ruminant, hindgut fermenters, and monogastric animals. Microbiologyopen. 2017;6:e00509.
22. Anand S, Kaur H, Mande SS. Comparative in silico analysis of butyrate production pathways in gut commensals and pathogens. Front Microbiol 7 (2016), doi:10.3389/fmicb.2016.01945.

23. Hungate RE, *The Rumen and its Microbes* (Elsevier Science, 1966).
24. Rémond D, Ortigues I, Jouany J-P, Energy substrates for the rumen epithelium. *Proc. Nutr. Soc.* **54**, 95–105 (1995).
25. Reichardt N, Duncan SH, Young P, Belenguer A, McWilliam Leitch C, Scott KP, Flint HJ, Louis P. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. ISME J. 2014;8:1323–35.
26. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, De Angelis M. The controversial role of human gut lachnospiraceae. Microorganisms 8 (2020),, doi:10.3390/microorganisms8040573.
27. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol. 2014;12:661–72.
28. Shetty SA, Marathe NP, Lanjekar V, Ranade D, Shouche YS. Comparative genome analysis of Megasphaera sp. reveals niche specialization and its potential role in the human gut. PLoS One. 2013;8:79353.
29. Makki K, Deehan EC, Walter J, Bäckhed F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. Cell Host Microbe. 2018;23:705–15.
30. Li J, Zhong H, Ramayo-Caldas Y, Terrapon N, Lombard V, Potocki-Veronese G, Estellé J, Popova M, Yang Z, Zhang H, Li F, Tang S, Yang F, Chen W, Chen B, Li J, Guo J, Martin C, Maguin E, Xu X, Yang H, Wang J, Madsen L, Kristiansen K, Henrissat B, Ehrlich SD, Morgavi DP. A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment. Gigascience. 2020;9:1–15.
31. Li J, Wang J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, Bork P. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41.
32. Ransom-Jones E, Jones DL, McCarthy AJ, McDonald JE. The Fibrobacteres: An Important Phylum of Cellulose-Degrading Bacteria. Microb Ecol. 2012;63:267–81.
33. Raut MP, Couto N, Karunakaran E, Biggs CA, Wright PC. Deciphering the unique cellulose degradation mechanism of the ruminal bacterium Fibrobacter succinogenes S85. Sci Rep. 2019;9:1–15.
34. Arntzen M, Várnai A, Mackie RI, Eijsink VGH, Pope PB. Outer membrane vesicles from Fibrobacter succinogenes S85 contain an array of carbohydrate-active enzymes with versatile polysaccharide-degrading capacity. Environ Microbiol. 2017;19:2701–14.
35. Terrapon N, Lombard V, Drula E, Lapébie P, Lapébie L, Al-Masaudi S, Gilbert HJ, Henrissat B. PULDB: the expanded database of Polysaccharide Utilization Loci. Nucleic Acids Res. 2017;46:677–83.
36. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. PLoS Biol 9 (2011), doi:10.1371/journal.pbio.1001221.

37. Tamura K, Hemsworth GR, Déjean G, Rogers TE, Pudlo NA, Urs K, Jain N, Davies GJ, Martens EC, Brumer H. Molecular Mechanism by which Prominent Human Gut Bacteroidetes Utilize Mixed-Linkage Beta-Glucans, Major Health-Promoting Cereal Polysaccharides. Cell Rep. 2017;21:417–30.
38. Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, Klinter S, Pudlo NA, Urs K, Koropatkin NM, Creagh AL, Haynes CA, Kelly AG, Cederholm SN, Davies GJ, Martens EC, Brumer H. A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. Nature. 2014;506:498–502.
39. Larsbrink J, Thompson AJ, Lundqvist M, Gardner JG, Davies GJ, Brumer H. A complex gene locus enables xyloglucan utilization in the model saprophyte Cellvibrio japonicus. Mol Microbiol. 2014;94:418–33.
40. Senoura T, Ito SS, Taguchi H, Higa M, Hamada S, Matsui H, Ozawa T, Jin S, Watanabe J, Wasaki J, Ito SS. New microbial mannan catabolic pathway that involves a novel mannosylglucose phosphorylase. Biochem Biophys Res Commun. 2011;408:701–6.
41. de Souza AP, Leite DCC, Pattathil S, Hahn MG, Buckeridge MS. Composition and Structure of Sugarcane Cell Wall Polysaccharides: Implications for Second-Generation Bioethanol Production. BioEnergy Res. 2013;6:564–79.
42. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.
43. Hemsworth GR, Thompson AJ, Stepper J, Sobala ŁF, Coyle T, Larsbrink J, Spadiut O, Goddard-Borger ED, Stubbs KA, Brumer H, Davies GJ. Structural dissection of a complex Bacteroides ovatus gene locus conferring xyloglucan metabolism in the human gut. Open Biol 6 (2016), doi:10.1098/rsob.160142.
44. Rohman A, Van Oosterwijk N, Puspaningsih NNT, Dijkstra BW. Structural basis of product inhibition by arabinose and xylose of the thermostable GH43 β-1,4-xylosidase from Geobacillus thermoleovorans IT-08. PLoS One 13 (2018), doi:10.1371/journal.pone.0196358.
45. Zanphorlin LM, Morais MAB, Diogo JA, Domingues MN, Souza FHM, Ruller R, Murakami MT. *Biotechnol. Bioeng.*, in press, doi:10.1002/bit.26899.
46. Grépinet O, Chebrou MC, Béguin P. Purification of Clostridium thermocellum xylanase Z expressed in Escherichia coli and identification of the corresponding product in the culture medium of C. thermocellum. J Bacteriol. 1988;170:4576–81.
47. Schubot FD, Kataeva IA, Blum DL, Shah AK, Ljungdahl LG, Rose JP, Wang BC. Structural basis for the substrate specificity of the feruloyl esterase domain of the cellulosomal xylanase Z from Clostridium thermocellum. Biochemistry. 2001;40:12524–32.
48. Pijning T, van Pouderoyen G, Kluskens L, van der Oost J, Dijkstra BW. The crystal structure of a hyperthermoactive exopolygalacturonase from *Thermotoga maritima* reveals a unique tetramer. FEBS Lett. 2009;583:3665–70.

49. Jung WS, Hong CK, Lee S, Kim CS, Kim SJ, Il Kim S, Rhee S. Structural and functional insights into intramolecular fructosyl transfer by inulin fructotransferase. J Biol Chem. 2007;282:8414–23.
50. Stender EGP, Andersen CD, Fredslund F, Holck J, Solberg A, Teze D, Peters GHJ, Christensen BE, Aachmann FL, Welner DH, Svensson B. Structural and functional aspects of mannuronic acid-specific PL6 alginate lyase from the human gut microbe Bacteroides cellulosilyticus. J Biol Chem. 2019;294:17915–30.
51. Boraston AB, Abbott DW. Structure of a pectin methylesterase from Yersinia enterocolitica. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2012;68:129–33.
52. Yu Z, Morrison M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. Biotechniques. 2004;36:808–12.
53. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R, Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4516–4522 (2011).
54. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
55. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013;10:996–8.
56. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res 42 (2014), doi:10.1093/nar/gkt1244.
57. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7 (2016), doi:10.1038/ncomms11257.
58. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
59. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
60. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605–7.
61. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
62. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
63. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H, Valencia A. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36:2251–2.

64. Prestat E, David MM, Hultman J, Taş N, Lamendella R, Dvornik J, Mackelprang R, Myrold DD, Jumpponen A, Tringe SG, Holman E, Mavromatis K, Jansson JK. FOAM (Functional Ontology Assignments for Metagenomes): A Hidden Markov Model (HMM) database with environmental focus. Nucleic Acids Res 42 (2014), doi:10.1093/nar/gku702.
65. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:D490-5.
66. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
67. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 2013;4:2304.
68. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz848.
69. Na S-Il, Kim YO, Yoon S-HH, min Ha S, Baek I, Chun J. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. J Microbiol 56 (2018), doi:10.1007/s12275-018-8014-6.
70. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30 (2013), doi:10.1093/molbev/mst010.
71. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
72. Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, Bokatzian S, Reisdorph N, Lozupone CA. AMON: Annotation of metabolite origins via networks to integrate microbiome and metabolome data. BMC Bioinformatics. 2019;20:614.
73. Santos CR, Costa PACR, Vieira PS, Gonzalez SET, Correa TLR, Lima EA, Mandelli F, Pirolla RAS, Domingues MN, Cabral L, Martins MP, Cordeiro RL, Junior AT, Souza BP, Prates ÉT, Gozzo FC, Persinoti GF, Skaf MS, Murakami MT. Structural insights into β-1,3-glucan cleavage by a glycoside hydrolase family. Nat Chem Biol. 2020. doi:10.1038/s41589-020-0554-5.
74. Mandelli F, De Morais MAB, De Lima EA, Oliveira L, Persinoti GF, Murakami MT. Spatially remote motifs cooperatively affect substrate preference of a ruminal GH26-type endo-β-1,4-mannanase. J Biol Chem. 2020;295:5012–21.
75. Kabsch W, B. A. T., D. K., K. P. A., D. K., M. S., R. R. B. G., E. P., F. S., W. K., K. W., K. W., K. W., K. W., K. P., W. M. S., *XDS*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 125–132 (2010).
76. Sheldrick GM. Experimental phasing with SHELXC/D/E: Combining chain tracing with density modification. Acta Crystallogr Sect D Biol Crystallogr. 2010;66:479–85.
77. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ, Adams PD, in *Acta Crystallographica Section D: Biological Crystallography* (International Union of Crystallography, 2007), vol. 64, pp. 61–69.
78. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS,

Terwilliger TC. P. H. Zwart, PHENIX: A comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr Sect D Biol Crystallogr. 2010;66:213–21.

79. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD. Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr Sect D Biol Crystallogr. 2012;68:352–67.
80. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr Sect D Biol Crystallogr. 1997;53:240–55.
81. Emsley P, Cowtan K. Coot: Model-building tools for molecular graphics. Acta Crystallogr Sect D Biol Crystallogr. 2004;60:2126–32.
82. Painter J, Merritt EA. TLSMD web server for the generation of multi-group TLS models. J Appl Crystallogr. 2006;39:109–11.
83. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC, MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 12–21 (2010).
84. Joosten RP, Long F, Murshudov GN, Perrakis A. The PDB-REDO server for macromolecular structure model optimization. IUCrJ. 2014;1:213–20.
85. Hammersley AP, Svensson SO, Hanfland M, Fitch AN, Häusermann D. Two-dimensional detector software: From real detector to idealised image or two-theta scan. High Press Res. 1996;14:235–48.
86. Svergun DI. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J Appl Crystallogr. 1992;25:495–503.
87. Svergun DI. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J. 1999;76:2879–86.
88. Volkov VV, Svergun DI. J Appl Crystallogr. 2003;36:860–4.
89. Kozin MB, Svergun DI. Automated matching of high- and low-resolution structural models. J Appl Crystallogr. 2001;34:33–41.
90. Miller GL. Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar. Anal Chem. 1959;31:426–8.
91. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr Sect D Biol Crystallogr. 2004;60:2256–68.
92. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41.

Figures

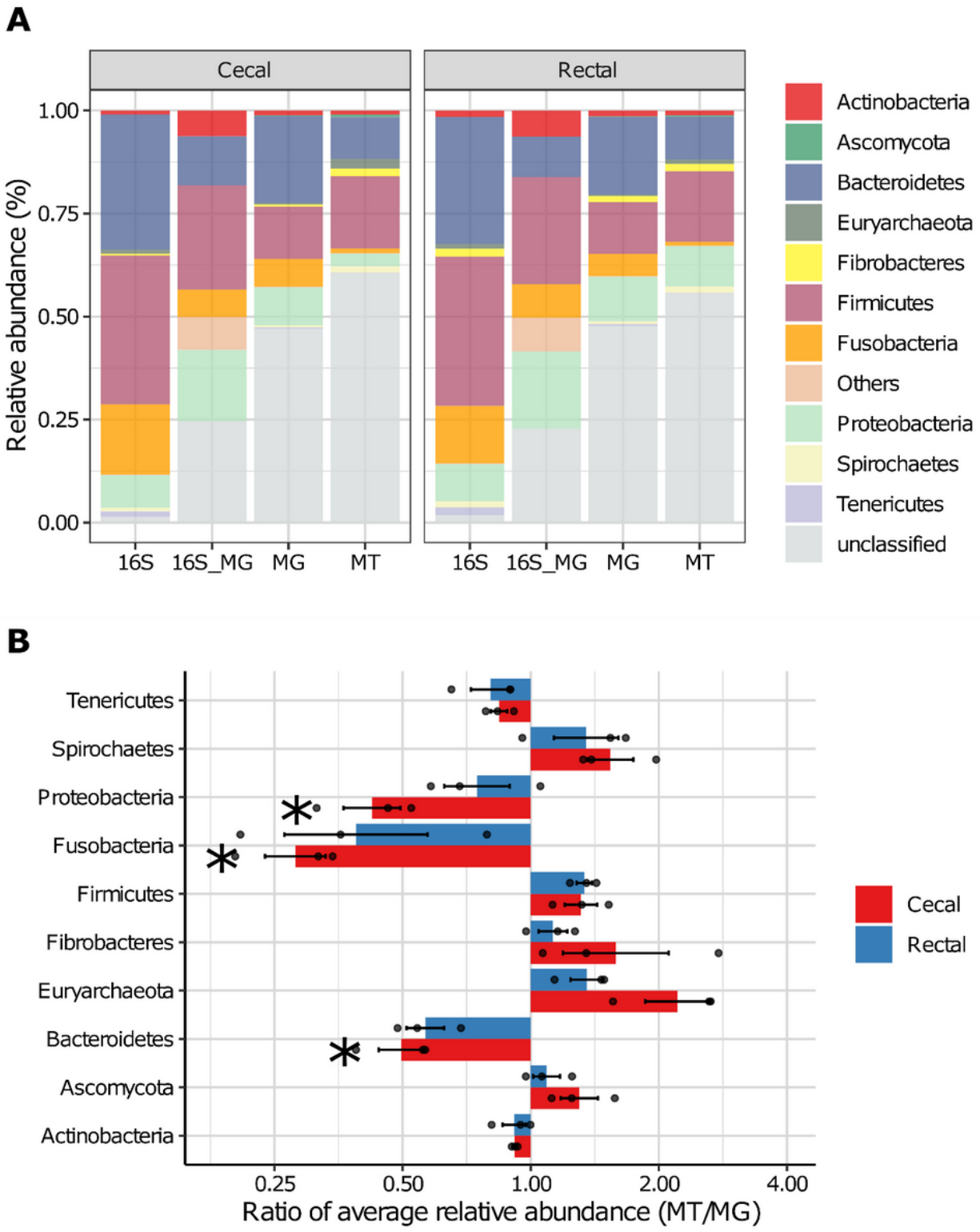

Figure 1

Microbial taxonomic composition of capybara gut microbiome. (A), Relative phyla abundance based on 16S rRNA target sequencing (16S), 16S rRNA recovered from metagenome (16S MG), whole metagenome (MG) and metatranscriptome (MT) reads. (B), Ratio of average relative phyla abundance of metatranscriptome to metagenome (MT/MG), expressed in log(2) scale. Asterisks represent significantly

different to MT/MG=1 (two-tailed t-test, FDR of 0.05 as threshold of significance level). The data represent the average of three independent experiments.

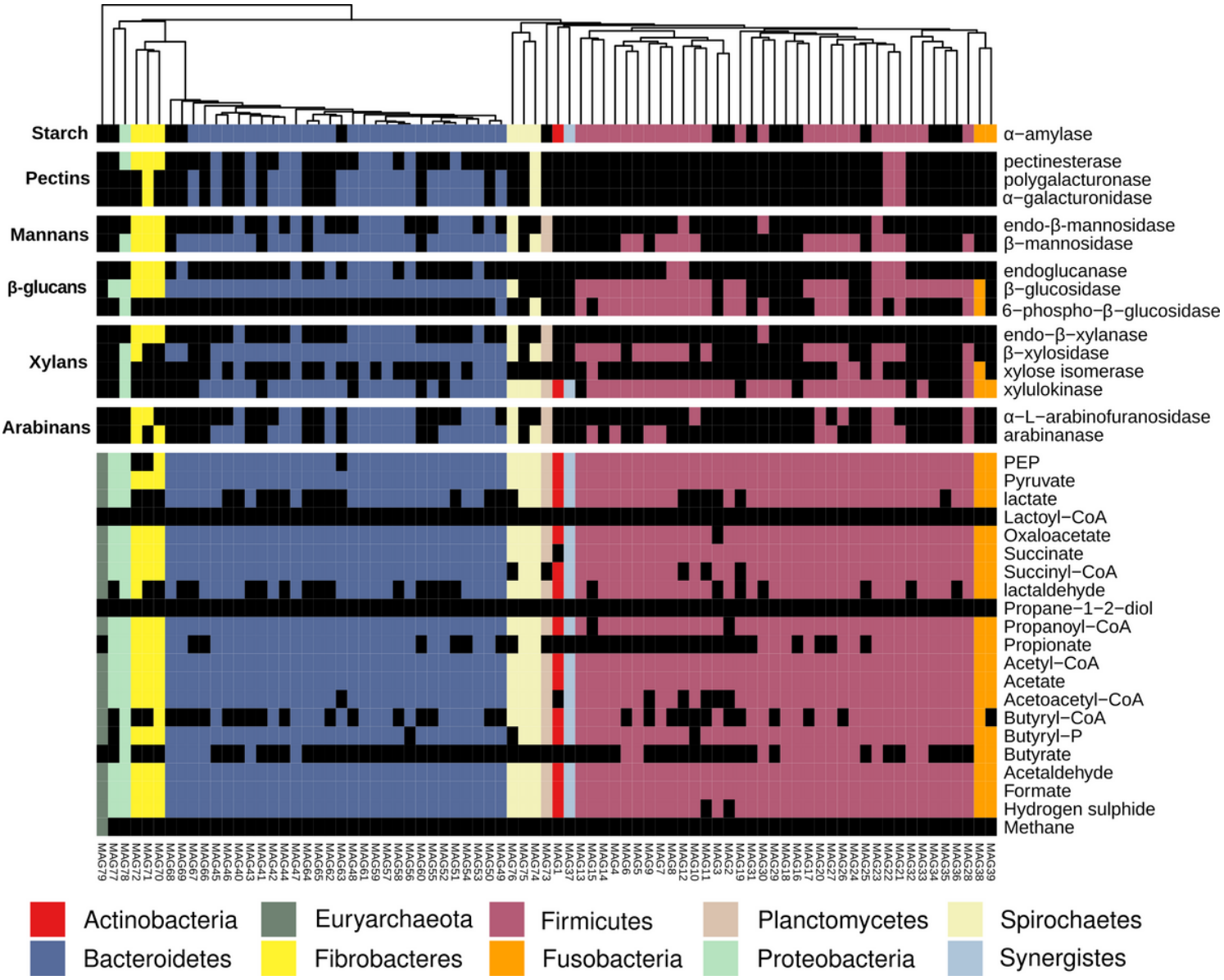

Figure 2

Metabolic reconstruction of 79 unique metagenome assembled genomes recovered from the capybara gut microbiome. Heatmap indicating presence or absence of enzymatic systems related to plant polysaccharides degradation or metabolites production (listed on the left) in each MAG (bottom) according to their set of genes. The presence of each enzyme/metabolite is denoted by a box colored by phylum taxonomy assignment, and black squares indicate the absence of the indicated metabolite/enzyme. Heatmap is clustered according to MAGs phylogeny.

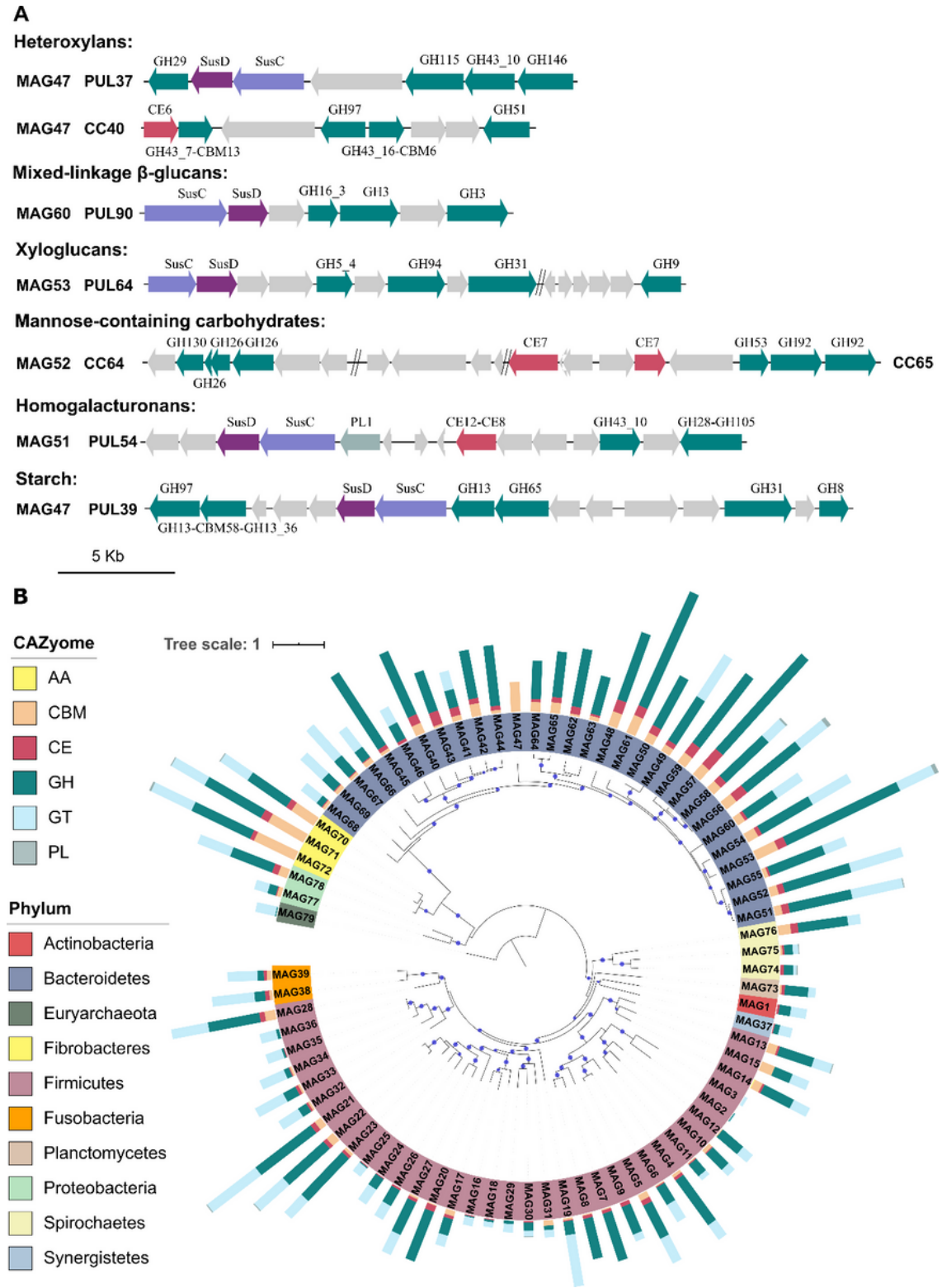

Figure 3

Main carbohydrate degradation systems from capybara gut metagenome assembled genomes. (A), Major PULs/CCs identified in the recovered genomes. Colored and gray arrows indicate identified and non-identified CAZyme sequences, respectively. (B), Maximum likelihood phylogeny of 79 unique MAGs reconstructed based on PhyloPhlAn pipeline(67). MAGs background is colored according to phylum taxonomy. Nodes with support values > 80 are indicated by a blue circle.

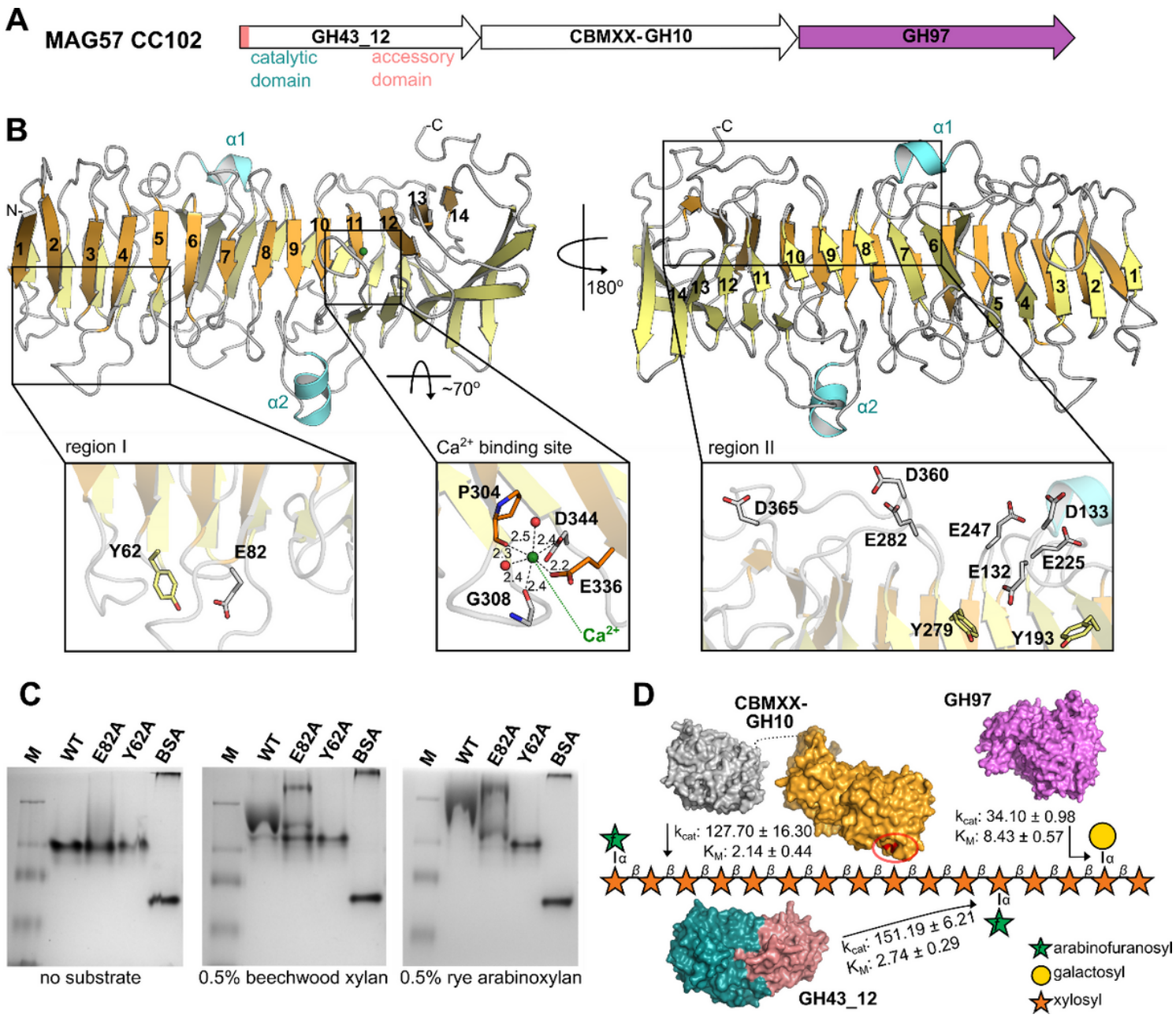

Figure 4

Enzymatic system for heteroxylan degradation from Prevotella sp. MAG57. (A), Schematic representation of the enzyme cluster (CC102) involved in heteroxylan breakdown. (B), Crystal structure of CapCBMXX indicating the β-helix fold consisting of 14 helical turns. The mutated residues from regions I and II, and the Ca2+ binding site are zoomed in (rectangles). (C), Affinity gel electrophoresis (AGE) with CapCBMXX (WT and mutants from region I) and xylan polysaccharides. Bovine serum albumin (BSA) was used as control (control) and M indicates the molecular weight markers. (D), Schematic representation of the modes of action of the enzymes CapCBMXX-GH10, CapGH97 and CapGH43_12 on heteroxylans. kcat values are expressed in s-1 and KM in mM (for CapGH97 and CapGH43_12) and mg.mL-1 for CapGH10. The residues Tyr62 and Glu82 are highlighted (red) in the CapCBMXX surface (orange) and the approximated possible binding region (region I) is indicated (red circle).

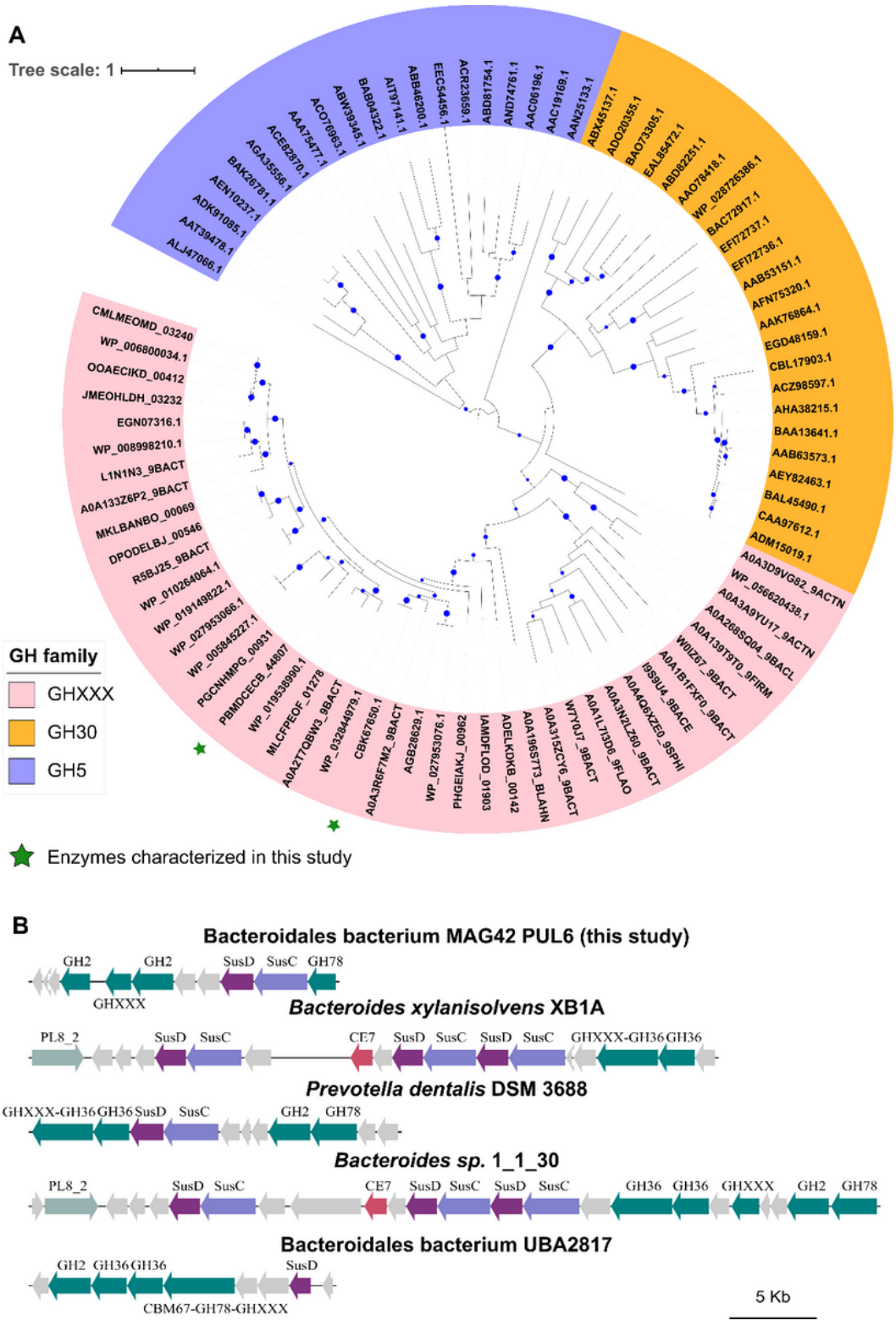

Figure 5

Phylogeny of the new GHXXX family and representative PULs. (A), Maximum likelihood phylogenetic analysis including GH30 (orange background) and GH5 (purple background) characterized members and proteins belonging to the new GHXXX family (pink background). Nodes with bootstrap support values > 50 are indicated by the blue circles. Founding members of GHXXX family characterized in this study are

denoted with a green star. (B), Representative GHXXX-containing PULs. Gray arrows indicate non-identified sequences.

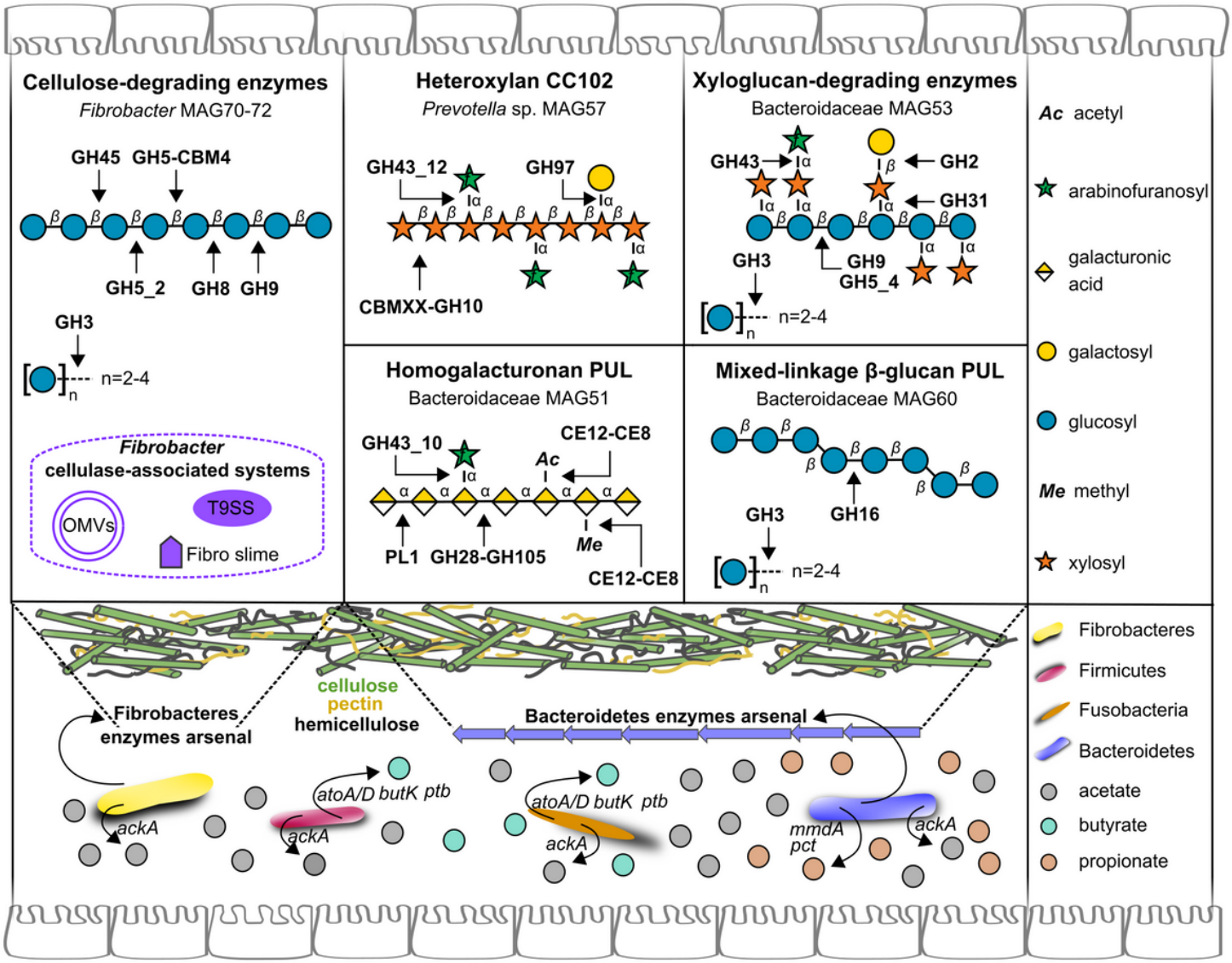

Figure 6

Schematic representation of the capybara gut microbial community and enzymatic strategies involved in the deconstruction and conversion of typical dietary polysaccharides into short chain fatty acids. In the upper panel is highlighted the CAZymes and mechanisms accounted to the depolymerization of cellulose and main hemicelluloses and pectins found in grasses such as sugarcane. In the lower panel is indicated the key phyla associated with hexoses and pentoses conversion into SCFAs. The upper and lower repeating drawings are representing the capybara intestinal villi.

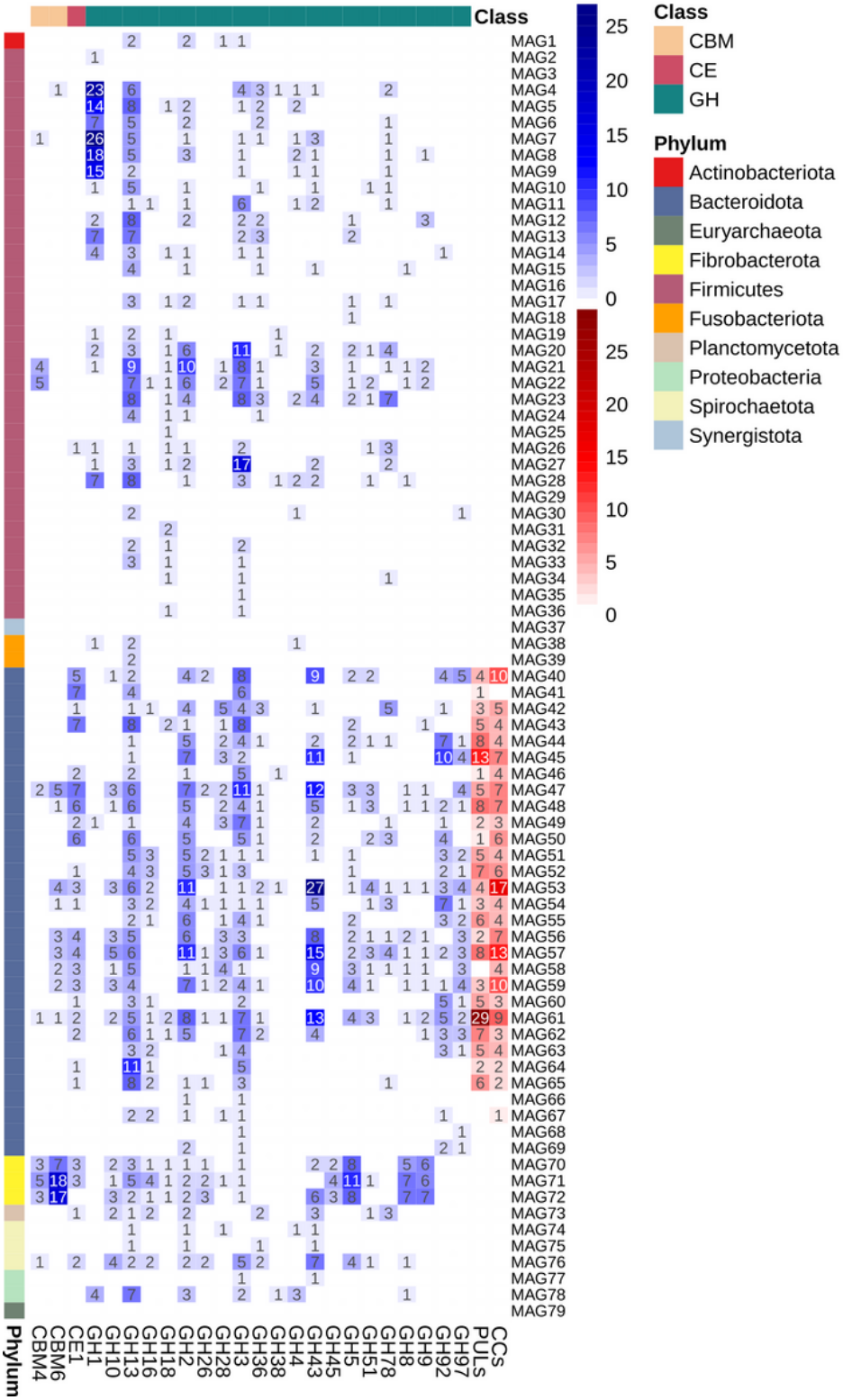

Figure 7

Heatmap of main CAZymes and PULs/CCs identified in the recovered genomes. The heatmap indicates the number of genes encoding CAZymes, PULs and CCs identified in each metagenome assembled genome. The number of PULs and CCs are colored in a red scale, whereas the number of each CAZy families are in a blue scale.

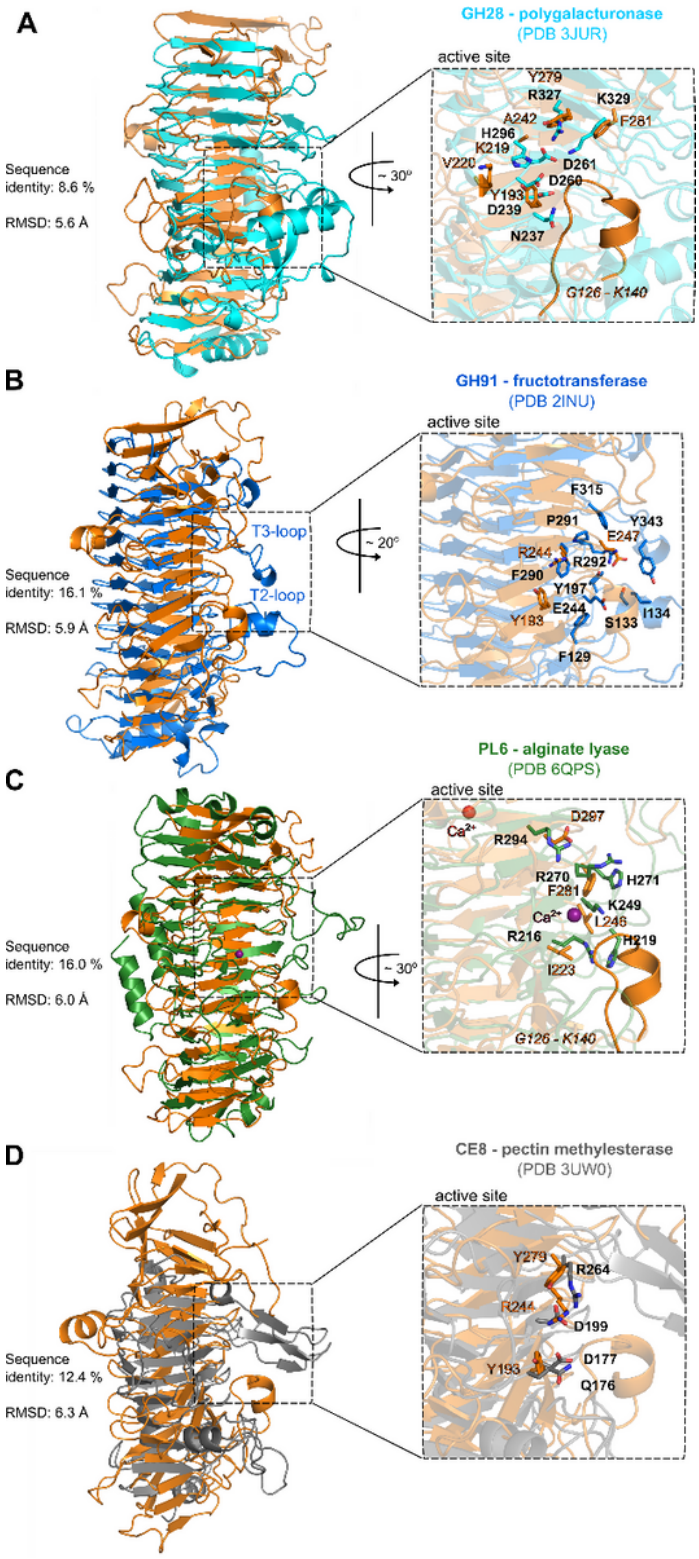

Figure 8

Structural comparisons of CBMXX with β-helix CAZyme families. A. Thermotoga maritima GH28 exo-polygalacturonase (PDB 3JUR) (cyan(48)). B. Bacillus sp. GH91 fructotransferase (PDB 2INU) (blue(49)). C. Bacteroides cellulosilyticus PL6 alginate lyase (PDB 6QPS) (green(50)) The Ca2+ ions are represented as red (CapCBMXX) and purple (alginate lyase) spheres. D. Yersinia enterocolitica pectin methylesterase (PDB 3UW0) (grey(51)). Structures chosen for comparisons with CapCBMXX were selected according to

higher Q-scores shown in Supplementary Table 10. Structural alignments and RMSD calculations were performed using Cealign plugin from Pymol and were adjusted to optimize the alignment of cavities (91). Sequence identities were obtained by pairwise alignment of PDB sequences (considering 1 protomer of each protein), using the EMBL-EBI search and sequence analysis tool (92).

Supplementary Files

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

- SuppTable8Crystallografic.pdf
- SupplementaryFigures.pdf
- SupplementaryTables17.xlsx
- SupplementaryTables913.xlsx