Université catholique de Louvain

Faculty of bioscience engineering

Bioprospecting for new carbohydrate active enzymes from the microbiomes of the termite gut and the anaerobic digestion process: an omics mediated approach

Marie BERTUCCI

Doctoral thesis in

Agricultural Sciences and Bioengineering

Committee:

Prof. Claude Bragard, UCLouvain (president)

Prof. Patrick GERIN, UCLouvain (supervisor)

Dr. Philippe DELFOSSE, Université du Luxembourg (supervisor)

Prof. Jacques MAHILLON, UCLouvain (secretary)

Dr. Magdalena CALUSINSKA, Luxembourg Institute of Science and Technology

Prof. Ralf KÔLLING, Universität Hohenheim

Abstract

Bioprospecting for new carbohydrate active enzymes from the microbiomes of the termite gut and the anaerobic digestion process: an omics mediated approach

Carbohydrate active enzymes play an important role in the biotechnological sector. They are key players in the deconstruction of complex polysaccharides, with applications for the production of bio-molecules and renewable fuels from lignocellulosic biomass. Discovering and increasing our knowledge on new carbohydrate active enzymes is crucial to improve the repertoire of available enzymes for the biorefinery industries.

In the context of bioprospecting for such enzymes, the aim of this PhD thesis was to characterize the carbohydrate hydrolytic potential of two lignocellulose-degrading microbiomes. I used metagenomics, metatranscriptomics and downstream bioinformatic analyses to investigate the polysaccharide deconstruction potential of microbiomes extracted from anaerobic digesters fed with sugar beet pulp, as well as from the gut of termites exposed to a strict miscanthus diet. I enlightened the genomic redundancy of the carbohydrate hydrolysis function: the genomic potential was maintained while the operational conditions and the related microbial community evolved. In the case of the termite's gut microbiome, an up regulation of carbohydrate active enzymes involved in lignocellulosic biomass deconstruction was observed. With the intention of confirming *in vitro* the substrate specificity predicted *in silico* for selected carbohydrate active enzymes identified in the metagenome, I developed a complete workflow, from DNA isolation to heterologous expression of the proteins (including proteins detection, proteins purification, and proteins identification) and activity detection.

This work led to the discovery of new carbohydrate active enzymes and related gene clusters, which brings additional knowledge to the scientific community towards the strategy employed by microorganisms to deconstruct lignocellulosic biomass. Additionally, these nature-optimised enzymatic cocktails might be mimicked and new specific enzymes be introduced in the precision biorefining pre-treatments of lignocellulosic biomass.

Acknowledgements

Firstly, I would like to express my sincere and immense gratitude to the best mentor I could never have imagined, Dr. Magdalena Calusinska. Few sentences will never be enough to thanks her for her continuous supervision and help during my PhD thesis. Besides being present, patient and always available for me, she demonstrated an inestimable support and kindness and I would never have been able to achieve this PhD without her and her immense knowledge.

I also would like to acknowledge my two supervisors, PhD Philippe Delfosse and Prof. Patrick Gerin, for their supervision and time, insightful comments and corrections as well as their encouragements during the whole duration of the thesis.

A special thanks goes to Prof. Claude Bragard, Prof. Jacques Mahillon and Prof. Ralf Kölling for their interest in my PhD thesis, their time spent reading my manuscript, their suggestions and corrections which helped me to improve this thesis.

Additionnally, I would like to thanks Dr. Corinne Rouland-Lefevre and Stephanie Giusti-Miller who gave me the opportunity to access their laboratory at the IRD Bondy, and for the time they spent giving me an intensive training which has been essential to this PhD thesis.

This experience would not have been the same without all my labmates and collegues, and I would like to thanks especially Xavier Goux,, Boris Untereiner, Dominika Klimek and Martyna Marynowska for their ideas, scientific and technical support and essentially for all the fun we had together during the last years.

Last but not least, my last acknowledgement goes to my parents, family and friends, and more particularly my boyfriend, who always believed in me and who spiritually supported me during this PhD thesis.

Table of contents

Abreviations

AA: auxiliary enzyme

ABC: ATP-binding cassette

AD: anaerobic digestion

ANI: average nucleotide identity

ARAF: α-L-arabinofuranosidases

AX: (arabino)xylans

bp: base pairs

CAZymes: carbohydrate active enzymes

CBM: carbohydrate-binding module

CE: carbohydrate esterase

CMC: carboxymethyl cellulose

COG: cluster of orthologous groups

EC: enzyme commission category

FA: ferulic acid

FAE: feruloyl esterase

gaxPUL: (glucurono)arabinoxylan targeting polysaccharide utilization loci

GH: glycoside hydrolase

gpPUL: Gram-positive polysaccharide utilization loci

GT: glycoside transferase

HRT: hydraulic retention time

IPTG: isopropyl-β-D-thiogalactopyranoside

KEGG: Kyoto Encyclopedia of Genes and Genome

KO: KEGG Orthology

LCB: lignocellulosic biomass

LPMO: lytic polysaccharide monooxygenase

MAG: metagenome assembled genome

Mbp: million base pairs

MFS: major facilitator superfamily

MG: metagenomics

MT: metatranscriptomics

Ni-NTA: nickel-nitrilotriacetic acid

OLR: organic loading rate

ORF: open reading frame

OTU: operational taxonomic units

PCR: polymerase chain reaction

PL: polysaccharide lyase

PTS: phosphotransferase system

PUL: polysaccharide utilization locus

SDS-PAGE: sodium dodecyl sulfate–polyacrylamide gel electrophoresis

RP: recombinant protein

RPKM: reads per kilobase million

SGBP: cell-surface glycan-binding protein

TBDT: TonB-dependent transporter

TPM: transcripts per kilobase million

VFA: volatile fatty acid

XG: xyloglucans

Definitions

16S analysis: Study of the bacterial 16S rRNA genes, revealing the bacterial community structure of the studied environment (microbiome).

(Metagenomic) Abundance: Number of reads that map (with >95% homology) to a given contig, normalized by the length of this contig (in kbp) and normalized again per million of mappable reads obtained in the same sample (total number of reads in the sample divided by 10^6). Abundance is expressed in Reads Per Kilobase Million (RPKM), i.e. reads per million of mapped reads and per kilobase of contig. Each gene encoded within the same contig has the same abundance.

Average abundance/expression: Mean of individual gene abundances/ expressions RPKM/TPM values for a set of genes.

Binning: Process aiming at grouping and assigning contigs to operational taxonomic units (group of closely related microorganisms).

Cluster of orthologous groups (COG) of proteins: Group of proteins found to be orthologous, that have typically the same function, and are present in different microbial lineages.

Contig: Assembled set of overlapping DNA sequences coding for one or more genes.

Cumulative abundance/expression: Sum of individual gene abundances/ expressions RPKM/TPM values for a set of genes.

Diversity of GHs/CAZys: Number of GH /CAZys families within the studied sample or metagenome assembled genome.

Gene expression: Number of mRNA reads that map (with >95% homology) to a given gene transcript, normalized by the length of this gene transcript (in kbp) and normalized again per million mappable reads obtained in the same sample (total number of mRNA reads in the sample divided by 10^6). The unit of gene expression is Transcripts per kilobase million (TPM), i.e. reads per million of mapped reads and per kilobase transcript.

Gene transcript: mRNA sequence.

Genes number: Number of nucleotide sequences identified as a gene, but with sequences differing by more than 5% (genes with bp sequences that have more than 95% homology are considered as one single gene).

High throughput sequencing: Automated techniques allowing the sequencing of hundred of millions of DNA molecules simultaneously and in a relatively short time, i.e. up to few days.

Holobiont: Ecological unit composed of an organism (the host) and the species living in (or around) the host.

Kyoto Encyclopedia of Genes and Genomes (KEGG): Database integrating genomic, systemic functional and chemical information relying on nucleotide sequence similarities.

Metagenome assembled genome (MAG): Genome reconstructed from metagenomic data analysis.

Metagenomics: High throughput analysis of the genomic content of the (micro)organisms present in the studied environment (microbiome).

Metatranscriptomics: High throughput analysis of the genomic expression (= mRNA) of (micro)organisms present in the studied environment (microbiome).

Microbiome: Community of microbial species present in a specific environment

OMICS analysis: High throughput analysis of DNA (genomics), mRNA (transcriptomics), proteins (proteomics) and/or metabolites (metabolomics) of a specific environment.

Reads: Nucleic acid sequences obtained from high throughput sequencing.

Relative (metagenomic /metatranscriptomic) abundance (%): Abundance of a gene/contig/MAG in the studied sample/environment divided by the total abundance of genes/contigs/MAGs identified in the same sample/environment; *e.g*. relative GH abundance: abundance of specific GH genes divided by the total abundance of all GH genes.

Chapter 1: General Introduction, objectives and approach

This thesis has been part of a core project funded by the FNR, called OPTILYS, which aims at (1) investigating the higher termite lignocellulotic system by combining metagenomics and metatranscriptomics approaches, (2) discovering new carbohydrate active enzymes and produce enzymatic cocktails towards the efficient degradation of lignocellulosic biomass. Therefore, two PhD theses were embedded in this project; one directed by Martyna Marynowska, aiming at investigating the higher termite lignocellulotic system, therefore, generating metagenomic and metatranscriptomic datasets. These datasets were further used for the second PhD thesis of the project (presented here), in order to discover new carbohydrate active enzymes and produce enzymatic cocktails. It is important to mention that the work done during this PhD thesis was also supported by two other FNR funded projects: GASPOP and LEGELIS, in order to investigate the microbiome of anaerobic digesters following metagenomics (GASPOP) combined with metatranscriptomics (LEGELIS) analyses. As for the OPTILYS project, omics datasets were not generated during the framework of this PhD thesis, but were used in order to assess the carbohydrate hydrolytic potential of anaerobic digesters and to find novel carbohydrate active enzymes.

1. Towards the use of lignocellulosic biomass

1.1. Environmental concern

Nowadays, an important concern is linked to environment. The humanity has been using natural resources faster than the planet's ecosystems is able to regenerate, and every year the Earth Oveshoot Day (i.e. the day when the amount of natural resources normally available for the year has already been used) comes earlier than the previous one (Figure 1).

The dependency on fossil fuels to meet the current energy needs will not be sustainable any longer, regarding the diminishing natural fuel reserves. The main negative effect related to this dependency is the impact on climate change through the emission of green house gases. Therefore, it is crucial that the humanity finds and makes use of alternative and renewable resources. In this sense, lignocellulosic biomass (LCB) is seen as an alternative source of biofuels, as well as a source of biopolymers for diverse industrial applications (Figure 2). In this PhD thesis, LCB mainly refers to agricultural residues, such as plant residues, rice or wheat straw as well as corn stover etc.

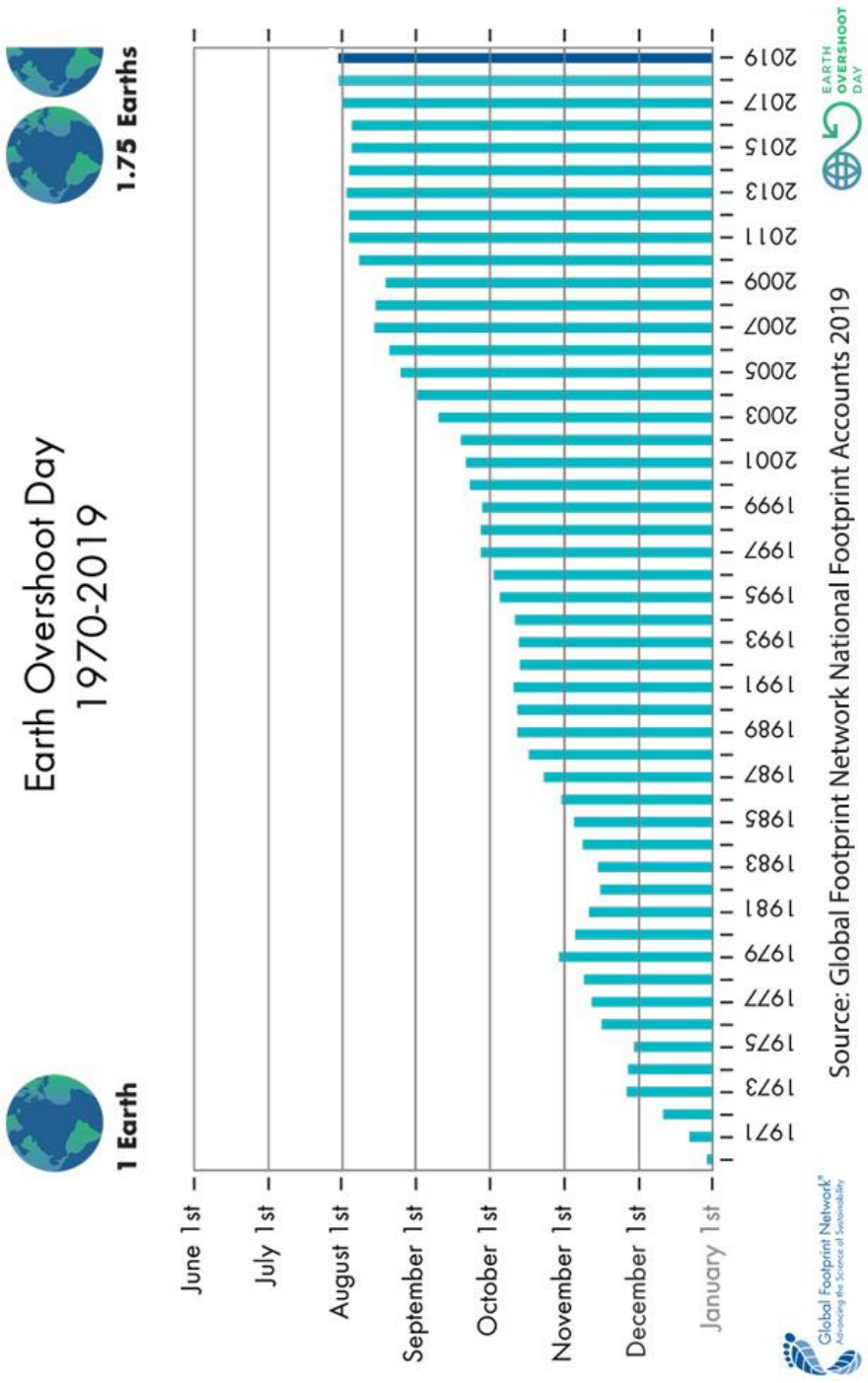


Figure 1: Representation of the Earth Overshoot Day from 1970 to 2019 (https://www.footprintnetwork.org).

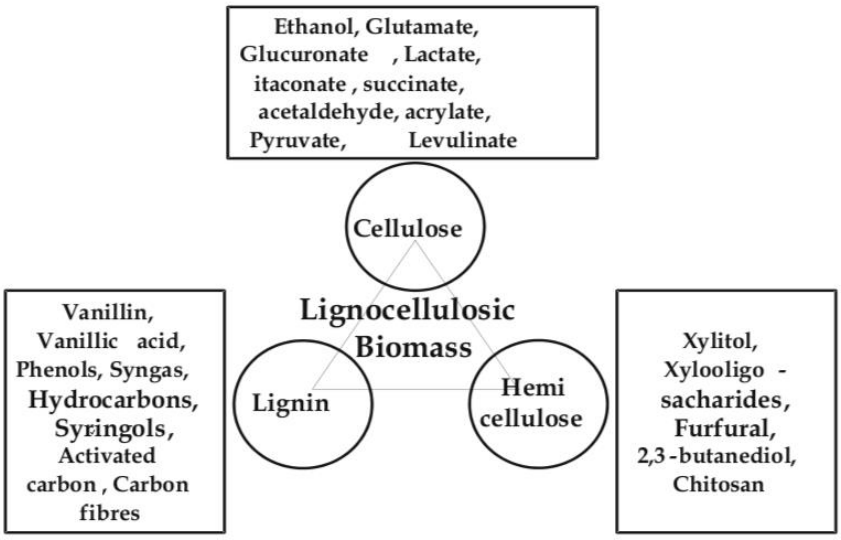


Figure 2: Common products that can be derived from lignocellulosic biomass components, adapted from Saini et al., 2019 (1)

1.2. Lignocellulosic biomass structure

Beside some exceptions (e.g. sugar beet pulp), LCB is mainly composed of cellulose, hemicellulose and lignin (Figure 2), and to a lower extend it contains small amounts of pectin, proteins and ash (2). However, depending on the type of LCB, the proportion of the different constituents is variable (Table 1). Additional factors such as plant age and climatic conditions, can affect LCB composition as well (3). The three major constituents, i.e. cellulose, hemicellulose and lignin, are intricately linked together through covalent and non-covalent bonds (4). Cellulose is a polysaccharide chain of glucose monomers interconnected through β-1,4 glycosidic linkages. These polysaccharidic chains are subsequently linked together via hydrogen bonds to form microfibrils (Figure 2) (5, 6). Hemicellulose corresponds to linear and/or branched chains of different sugar monomers linked through β-1,3 and/or β-1,4 glycosidic bonds. Hemicellulose polysaccharidic composition is variable depending on the LCB source (7). Nevetheless, the most common hemicellulosic polysaccharides include: xylan, mannan (and their derivatives), as well as xyloglucan and β-glucan. Xylan is often substitited with side-chains of glucuronic acid and/or arabinose moieties, and is thus called glucurono- and/or arabino-xylan. Similarly, galactomannan comprises

galactose residues branched to the mannan main chain. However, glucomannan is composed of mannose and glucose monomers in the main chain. Finally, lignin, the third most abundant component of LCB, it is a complex three-dimensional polymer composed of three phenolic building blocks (G-guaiacyl, S-syringyl and H-p-hydroxyphenyl). It is often ester-linked to hemicellulose thus rendering the biomass recalcitrant (resistant) to hydrolytic lysis.

Table 1: Composition of diverse lignocellulosic biomasses, adapted from Rezic et al., 2013; Yadav et al., 2018 (8, 9)

LCB	Cellulose (%)	Hemicellulose (%)	Lignin (%)
Rice straw	15-29	9-17	12-18
Wheat straw	34-39	21-34	22-25
Switch grass	31-38	26-34	18-22
Sugarcane bagasse	42-45	25-28	20
Miscanthus	20-40	23-35	19-31
Sugar beet pulp*	22-30	24-32	1

*Pectin composition: 38-62%
*LCB: Lignocellulosic biomass

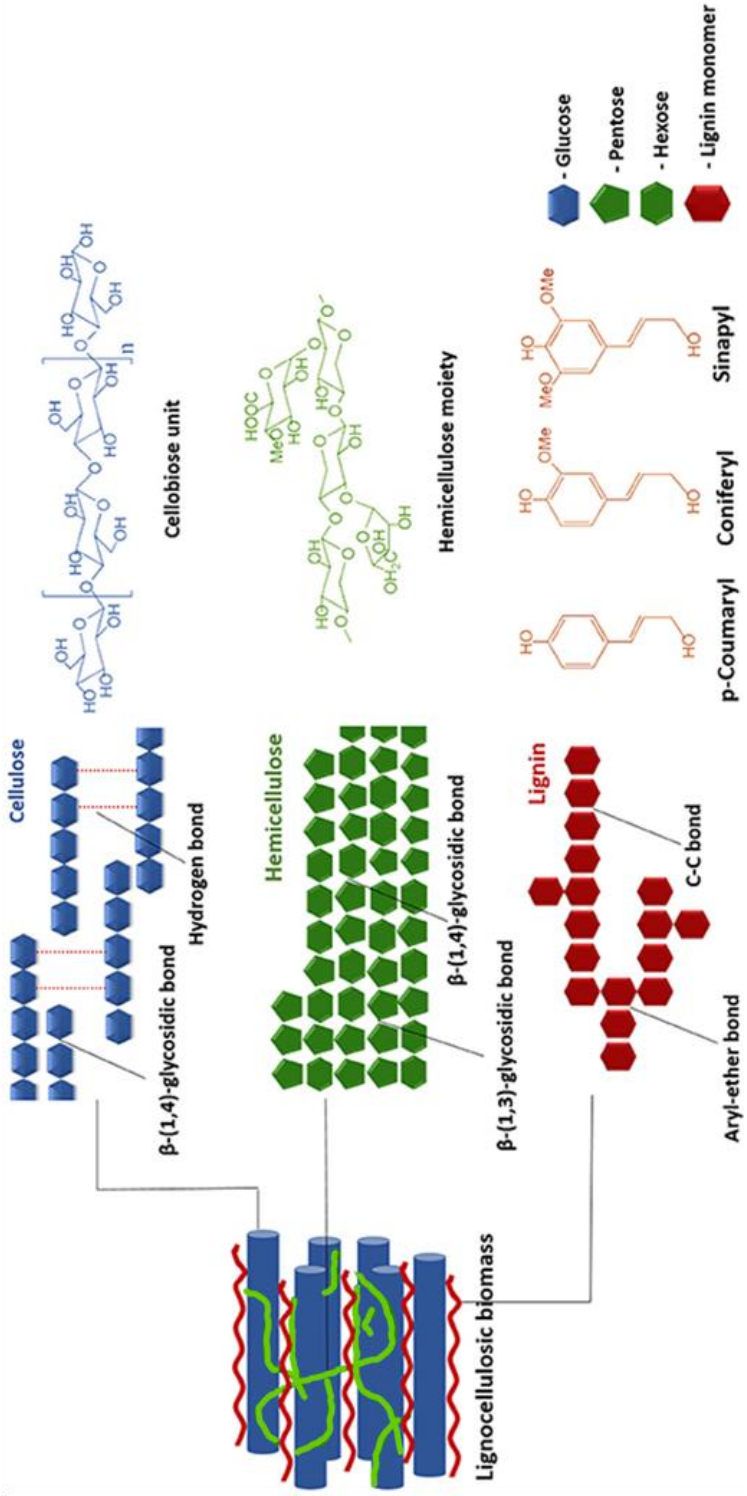


Figure 2: Structure of lignocellulose, adapted from Baruah et al., 2018 (5).

1.3. Current pre-treatments of lignocellulosic biomass

Due to its high recalcitrance and complex strucure, the deconstruction of LCB to monosugars is a multi-step mechanism. LCB resistance to enzymatic hydrolysis is often related to the crystallinity of cellulose in microfibrils which are coated with hemicellulose and together embedded by lignin polymers (10). Therefore, an efficient pre-treatment of LCB is required in order to detach cellulose and hemicellulose fibres from the lignin framework, before they could further be degraded into simpler sugars. In biorefineries, different pre-treatment strategies including enzymatic, thermal, freeze/thaw, chemical, wet oxidation etc. and their combination (11), have been developed to dissociate, modify or partially degrade lignin. Nevertheless, up to now, all the existing schemes suffer from diverse, inherent drawbacks, including among others large energy input and chemicals requirement (e.g. thermochemical pre-treatment), and the lack of specificity leading to the formation of unwanted compounds, such as inhibitors and pollutants, (e.g. furfurals and phenols resulting from acid and alkaline pre-treatments) (4).

Alternatively, biological pre-treatment, including the application of enzymes extracted from white rot fungi, has also been proposed and recently it has gained a lot of attention from the biorefinery industries. The biological approach has several advantages over thermochemical pre-treatments. Firstly, it can be applied at ambient or moderately warm temperatures, thus reducing thermal and electrical energy inputs. And secondly, it does not lead to the high production of harmful chemical by-products, which are known to inhibit further cellulose hydrolysis. Accellerase® TRIO™ (Genecor) is a commercial, enzymatic cocktail, produced using a genetically modified fungus *Trichoderma reesei*, and designed for an efficient hydrolysis of lignocellulosic biomass into fermentable monosaccharides. Interesting results were also reported for other fungal strains, including *Coriolus versicolor* (37% increase in the saccharification rate of pre-treated bamboo residues (12)), and *Ceriporiopsis subvermispora* (31% reduction in lignin content with only 6% loss in cellulose (13)). Most of these enzymatic cocktails, in some cases commercially available, are composed of different enzymes mixes, usually of a single organism origin, selected based on their complementary activities, sometimes mixed randomly (14). A speculative justification for using enzymes from the same organism is that they are more likely to be compatible with each other, because they co-evolved together.

2. Lignocellulosic biomass deconstruction by natural systems

2.1. The role of carbohydrate active enzymes in lignocellulosic biomass deconstruction

In living organisms, the deconstruction of cellulose and hemicellulose is mainly driven by the action of so called carbohydrate active enyzmes (CAZymes). They are responsible for the buiding, degradation and modification of carbohydrates and glycoconjugates (15). Depending on their mechanism, they can be classified into different classes, known as glycoside hydrolases (GHs), glycoside transferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), auxiliary enzymes (AAs) and carbohydrate-binding modules (CBMs). Inside the different categories, enzymes are further classified into families based on their sequence similarties. Currently, more than one million six hundred thousand proteins have been reported within the CAZy database (http://cazy.org). Table 2 reports on the number of families and proteins classified within each category. Additionally, enzymes can be numerically classified based on the chemical reaction they catalyze, to this purpose they are assigned to an enzyme commission category (EC).

Table 2: Inventory of the reported CAZy families and proteins referenced in the public database (http://cazy.org).

CAZy class	Number of assigned families	Number of classified proteins
GHs	164	704646
GTs	109	599236
PLs	38	21913
CEs	16	72102
AAs	16	12749
CBMs	86	189702

Cellulose and hemicellulose enzymatic deconstruction requires an arsenal of hydrolytic CAZymes, mainly representing GHs and CEs families, widespread within the CAZy database (Table 3). Due to its non polysaccharidic structure, lignin deconstruction is not driven by the action of CAZymes, but mainly be oxidoreductases (16). More specifically, lignin enzymatic deconstruction is performed by the combined action of laccases and diverses peroxidases, *e.g.* lignin peroxidase or manganese peroxidase (17).

Table 3: Diversity of enzymes within the CAZy database involved in cellulose and hemicellulose deconstruction (http://cazy.org)

Enzyme activity	**EC classification**	**CAZy family**
Endo-β-1,4-glucanase	3.2.1.4	GH5, 6, 7, 8, 9, 10, 12, 44, 45, 48, 51, 74, 124
Exo-β-1,4-glucanase	3.2.1.91 3.2.1.176	GH5, 6, 7, 9, 48
β-glucosidase	3.2.1.21	GH1, 3, 5, 9, 30, 116
Endo-β-1,4-xylanase	3.2.1.8	GH5, 8, 10, 11, 30, 43, 51, 98, 141
β-xylosidase	3.2.1.37	GH1, 2, 3, 30**,** 39, 43, 51, 52, 54, 116, 120
Acetylxylan esterase	3.1.1.72	CE1, 2, 3, 4, 5, 6, 7, 12
Feruloyl esterase	3.1.1.73	CE1
α-glucuronidase	3.2.1.139	GH4, 67
α-L-arabinofuranosidase	3.2.1.55	GH2, 3, 43, 51, 54, 62
endo-β-1,4-mannanase	3.2.1.78	GH5, 26, 45, 113, 134
β-mannosidase	3.2.1.25	GH1, 2, 5, 164
α-galactosidase	3.2.1.22	GH4, 27, 31, 36, 57, 97, 110

More specifically, the classic cellulose deconstruction is performed by the action of so called cellulases (Figure 4) comprising (1) endo-β-1,4-glucanase, i.e. EC 3.2.4.1, hydrolyzing β-1,4 bonds inside cellulose and releasing cellobiose, (2) exo-β-1,4-glucanases, i.e. cellobiohydrolase I EC 3.2.1.176, and cellobiohydrolase II EC 3.2.1.91, respectively acting on reducing ends and non-reducing ends and releasing cellobiose and cello-oligosaccharides, and (3) β-glucosidase, EC 3.2.1.21, degrading cellobiose and cello-oligosccharides to glucose. Cellulases catalytic domains are often associated with non-catalytic domains, i.e. CBMs, which are thought to enhance the enzymatic association with the substrate (18). More recently, it has been evidenced that different mechanisms to deconstruct cellulose are employed by some aerobic organisms (e.g. *basidiomycetes* and *ascomycetes* fungi (19) as well as *actinomycetes* bacteria (20, 21)). These mechanisms implicate the use of lytic polysaccharide monooxygenases (LPMOs). Through an oxidative mechanism, these enzymes are able to cleave the glycosidic bonds within cellulose (22).

The enzymatic hydrolysis of hemicellulose requires different sets of enzymes depending on the hemicellulosic structure. As an example, arabinoxylan and galactoglucomannan enzymatic degradation reactions are reprensented in Figure 5. The xylan backbone is depolymerized by the action of endo-β-1,4-xylanases, EC 3.2.1.8, which randomly attack the xylan within the main chain, releasing xylo-oligosaccharides. Furthermore, β-xylosidases, EC 3.2.1.37, attack the extremity of the xylan chain as well as they target xylobiose to liberate xylose. Additionally, side-chain residues are removed by the combined action of acetylxylan esterases, EC 3.1.1.72, feruloyl esterases, EC 3.1.1.73, α-glucuronidases, EC 3.2.1.139 and α-L-arabinofuranosidases, EC 3.2.1.55. Regarding galactoglucomannan deconstruction, usually it requires three types of enzymes including, endo-β-1,4-mannanases, EC 3.2.1.78, β-mannosidases, EC 3.2.1.25 as well as α-galactosidase, EC 3.2.1.22. Respectively, they are responsible for the hydrolysis of internal glycosidic bonds within the mannan backbone, the removal of mannose or glucose from the end-chain and from small oligosaccharides, and the hydrolysis of the galactose side-chain residue.

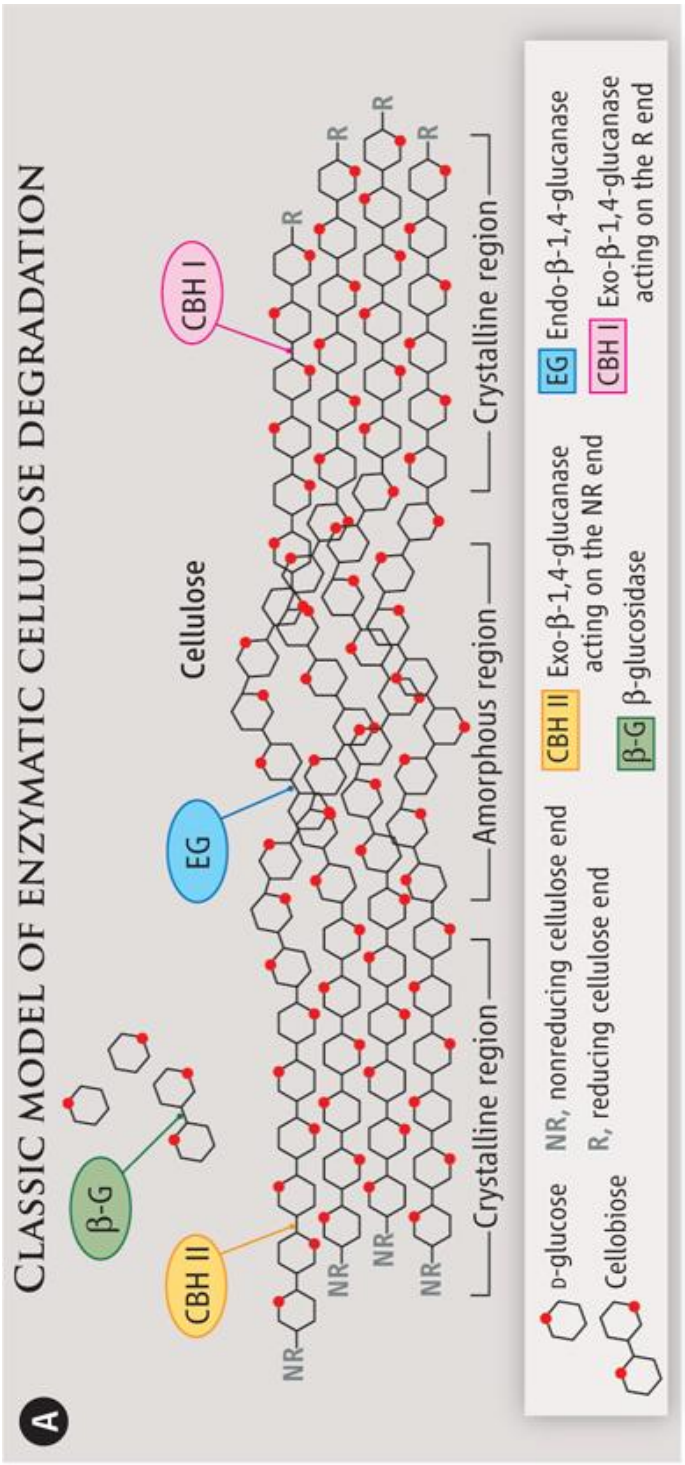


Figure 4: Classic model of cellulose deconstruction by carbohydrate active enzymes, adapted from Berlin, 2013 (98).

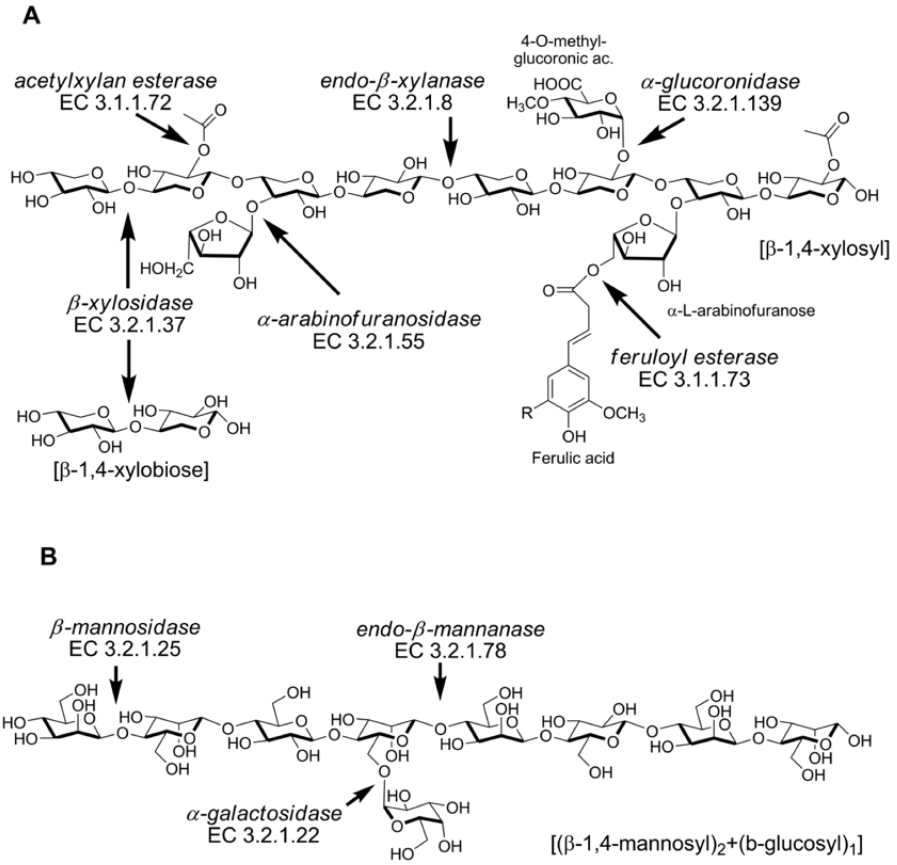


Figure 5: Arabinoxylan and galactoglucomannan enzymatic deconstruction, adapted from Linares-Pasten et al., 2014 (23)

2.2. Strategies employed by organisms to deconstruct lignocellulosic biomass

The strategies employed by different organisms to deconstruct LCB have been widely studied by the scientific community. Prokaryotes (mainly bacteria), together with eukaryotes (protozoa and fungi) harbor in their genomes CAZymes, pointing out to their ability to participate to LCB deconstruction (24). In ruminants and insects, LCB deconstruction rely on the symbiosis between the host and its microbiota which comprises bacteria, fungi and protozoa. It is worth mentioning that archaea and ciliate protozoa are engaged in a symbiotic relationship (25), however their role is mainly limited to the methanogenesis (26). Their implication in LCB deconstruction is less studied, however non methagenomic archaea can also deconstruct LCB, as evidenced by a consortium of three hyperthermophilic archeae (27).

In bacteria, CAZymes can be encoded separately in their genome or they can form cluster of colocalized genes, e.g. known as **polysaccharide utilization loci** (PULs). CAZymes can also be part of enzymatic structures such as cellulosomes. Cellulosomes have been largely invertigated in the past and are mainly found in the genomes of anaerobic bacteria, such as *Firmicutes* (particularly *Clostridia*). They were isolated from different anaerobic environments, e.g. soil, termite gut, anaerobic reactor, etc (28). They consist of a scaffolding backbone attached to the cell surface and harbouring several cohesion domains interacting with dockerin domains (Figure 6). The latest is mainly associated with mainly cellulases and in some cases also hemicellulases (29). Additionally, scaffoldins can contain CBMs that bind the enzymatic complex onto cellulose.

Independently, *Bacteroidetes* have developed their own polysaccharide deconstruction strategy: the PUL system. *Bacteroidetes* are gram negative bacteria, which are present in many different environments (e.g. anaerobic reactors, soil, gastrointestinal tract of animals, etc. (30)). The study of these environments led to the discovery of a unique strategy employed by *Bacteroidetes* to deconstruct polysaccharides. It was firstly evidenced by Anderson and Salyers in 1998 when studying the human gut microbiota (31). However, only recently the term PUL was introduced to the scientific literature (32), when the approach employed by

Bacteroidetes to deconstruct glycans was characterised. PULs are defined as clusters of genes, being colocalized as well as coregulated and being responsible for the detection, binding, digestion of complex polysaccharides and transport of hydrolysis products. Therefore, PULs contain a complete set of enzymes to target diverse polysaccharides. In this sense, nature-optimised PULs, encoded in *Bacteroidetes* genomes and targeting different LCB components, might be a promising solution for the emerging biorefinery sector. To be annotated as a PUL, the gene cluster needs to encode for a cell surface glycan-binding protein (i.e. namely SusD like gene) as well as a TonB-dependent transporter (i.e. namely SusC like gene). Accordingly to their names, these proteins are involved in the binding and the transport of polysaccharides (33). One example is the utilization of xyloglucan by *Bacteroides ovatus* in the human gut (Figure 7). Few other PULs have been discovered and identified, e.g. one targeting pectin (34) or galactomannan (35). Most commonly, PULs contain genes encoding for sensor-regulator systems bound to the inner membrane and controlling the expression of genes associated with the PUL. These sensor-regulators include : the hybrid two component system (HTCS), the SusR sensor/regulator or the extracytoplasmic function sigma/anti-sigma factor (ECF) (36). Even though oligosaccharides are known to activate the sensor-regulators, little is known about the actual molecular cues recognized (37).

Regarding the last major component of LCB, only some fungi and bacteria are able to deconstruct lignin, and its biodegradation has been widely studied within white-rot and brown-rot fungi (38, 39). Indeed, white-rot fungi are highly effective in lignocellulose deconstruction by producing extracellular enzymes directly involved in the deconstruction of the three main components of LCB (40–42). However, the process of lignin deconstruction by brown-rot fungi mainly involves non-enzymatic oxidation reactions, and therefore they are less efficient than white-rot fungi (17). On the other hand, research on bacterial lignin deconstruction is less advanced, but it shows that *Actinomycetes*, *α-proteobacteria* and *γ-proteobacteria* are able to solubilize polymeric lignin (43).

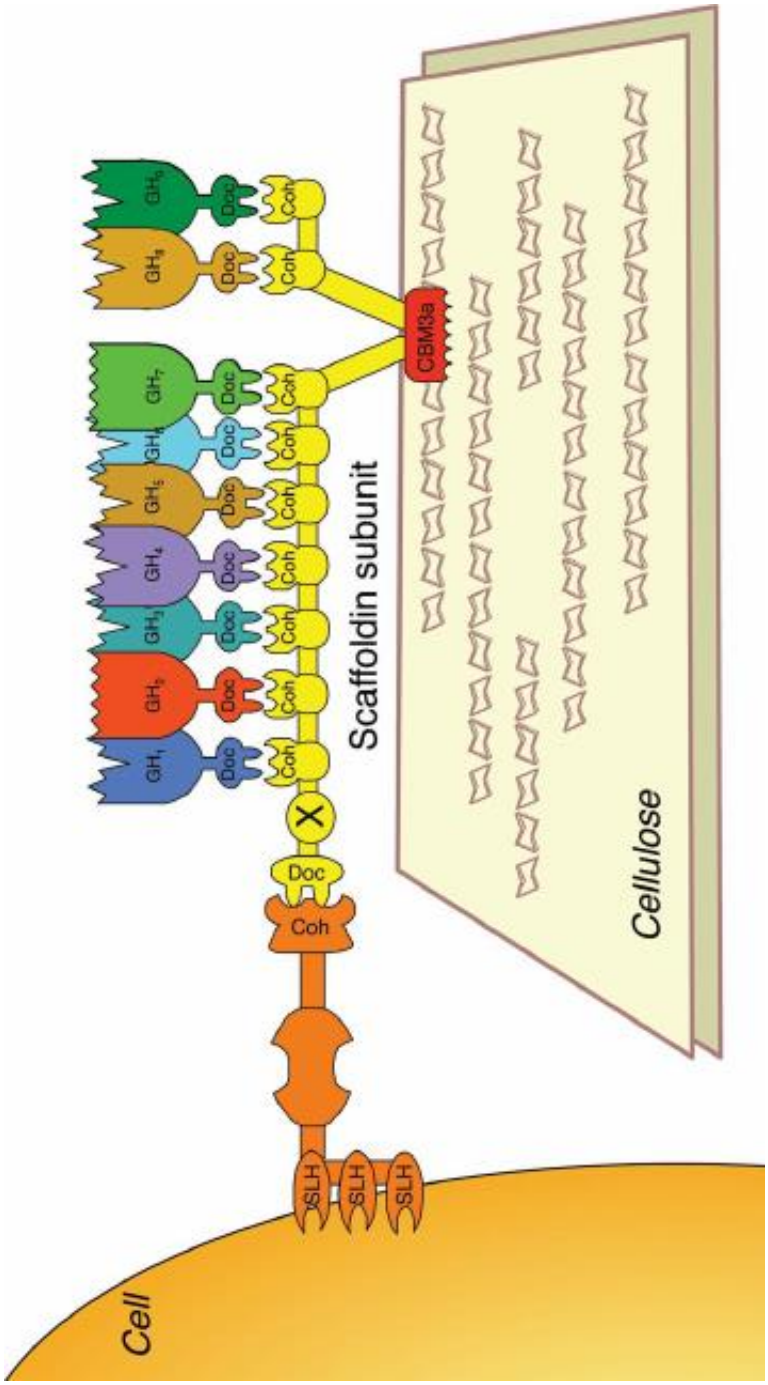


Figure 6: Schematic representation of the most studied and best characterized cellulosome from *Firmicutes* (*Clostridium thermocellum*) targeting cellulose deconstruction. SHL: surface layer homology like module attaching the scaffolding subunit to the cell wall, Coh: cohesins, Doc: dockerins, GH: glycoside hydrolase, CBM: carbohydrate binding modules, adapted from Yaniv et al., 2014 (99)

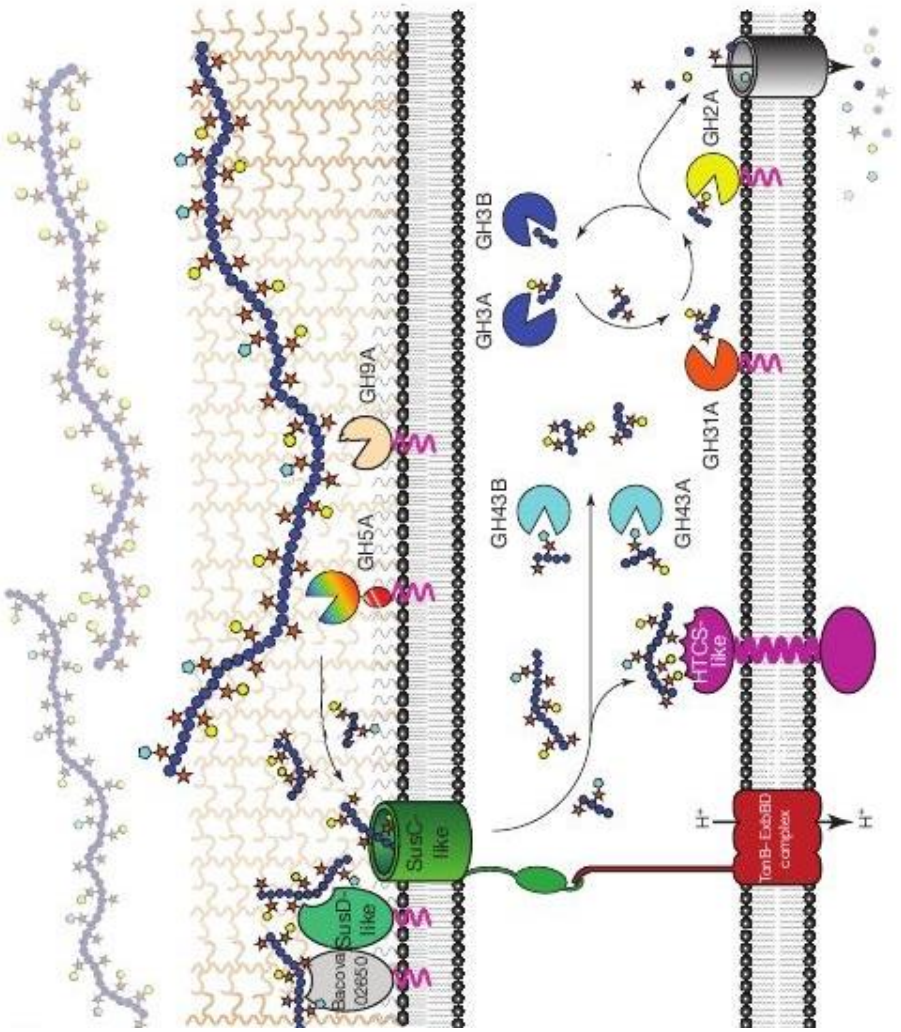


Figure 7: Representation of a model *B. ovatus* mechanism for the deconstruction of xyloglucan. Enzymes are represented as circles, colour-coded as follow: rainbow, endo-xyloglucanase BoGH5A; tan, endo-xyloglucanase BoGH9A; orange, α-xylosidase BoGH31A; turquoise, α-L-arabinofuranosides BoGH43A and/or BoGH43B; yellow, β-galactosidase BoGH2; dark blue, β-glucosidases BoGH3A and/or BoGH3B, adapted from Larsbrink et al., 2014 (44)

2.2.1. Lignocellulosic biomass deconstruction by termites

Across the tree of life, termites (the termite gut system) are considered the most efficient natural lignocellulose degraders, and they have been abundantly studied over the last years. Indeed, these insects developed a symbiotic system (enzymes of termite, bacterial, fungal or protozoal origins) that is more efficient than any known fungi, and that can degrade lignocellulose with an effectiveness ranging from 50 to almost 100% within less than 24h (45). In the termite gut system, mechanical deconstruction works together with enzymes to deconstruct lignocellulose. Additionally, lower termites (mostly wood-feeding) rely on the symbiotic action of flagellate protists and prokaryotes (46). Interestingly, higher termites lost protozoans, relying only on prokaryotes (Figure 8). Therefore, they developed new strategies to digest lignocellulose, such as association with fungus (*Basidiomycete* fungi-growing *Macrotermitinae* termites)(47), as well as adapted feeding habits (*e.g.* soil, dry grass, wood, litter)(45).

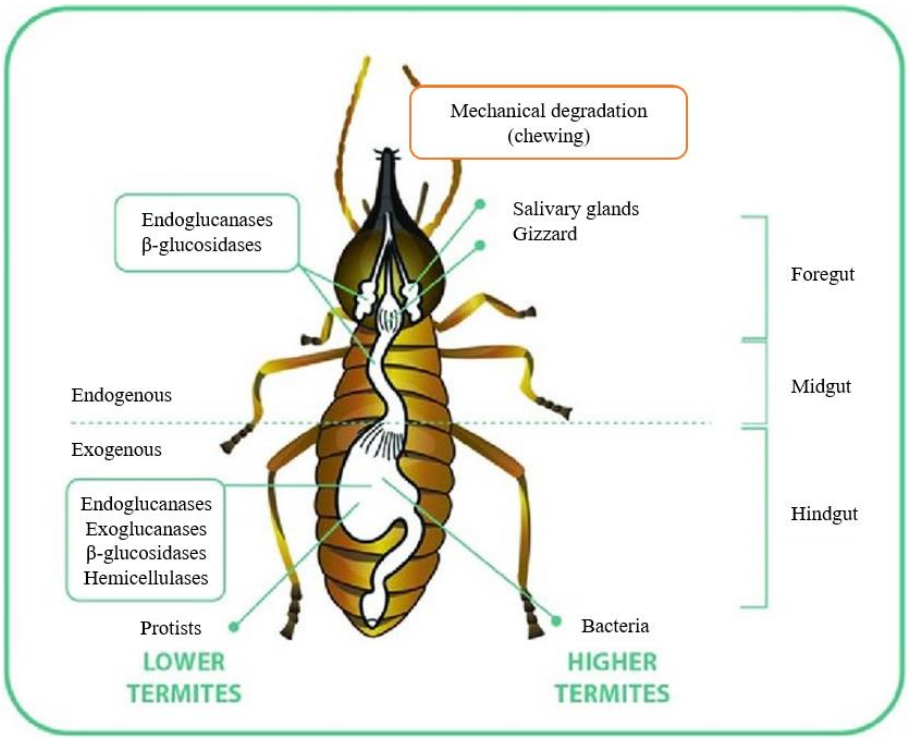


Figure 8: The digestive process of lignocellulose in termites, adapted from Talia et al., 2018 (48).

2.2.2. Lignocellulosic biomass deconstruction through anaerobic digestion

Next to natural processes, man-optimised processes have been developed in order to degrade and make use of LCB. One example is the anaerobic digestion (AD), an oxygen-free process of biomass deconstruction leading to the production of biogas (a mixture of methane, carbon dioxide and trace gases). AD occurs in natural ecosystems, including the digestive tract of animals e.g. cows, kangaroos, sheeps and deers, etc. (49).

AD is a multi-stage process divided into four main steps (Figure 9). Biomass enzymatic deconstruction occurs during the fist stage of the process, namely the hydrolysis. In this step, the microbial consortia present in the reactor degrade the complex organic matter, including polysaccharides, proteins and lipids into soluble organic molecules, i.e. monosugars, amino and fatty acids. During the next steps, i.e. acidogenesis, acetogenesis and methanogenesis, the soluble organic molecules are converted to the final product, the biogas. As such, anaerobic reactors are reservoirs of CAZymes and have been widely studied over the past decades.

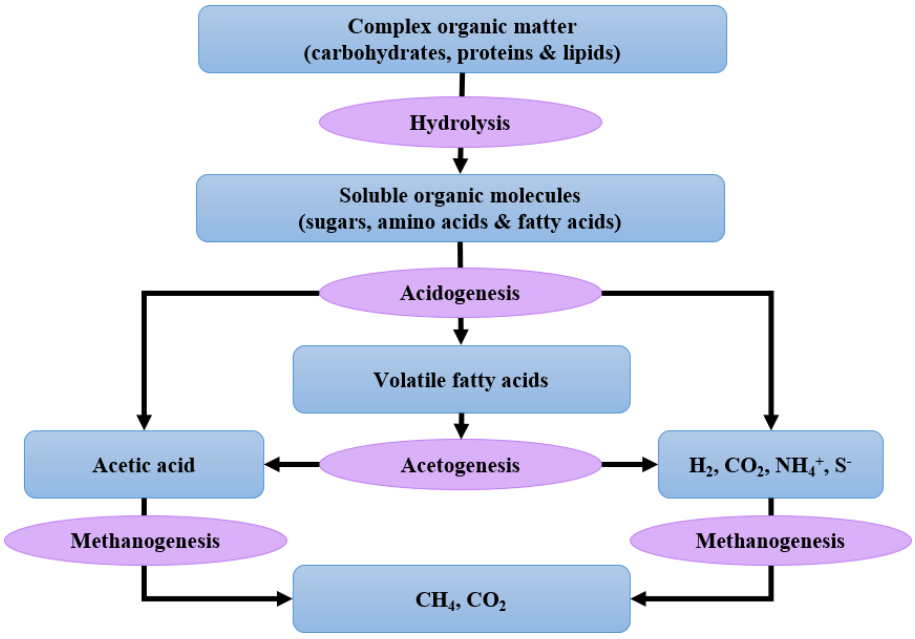


Figure 9: Steps involved in the anaerobic digestion process of LCB, adapted from Ramaraj et al., 2015 (50)

3. Omic-assisted technologies towards the characterisation of microbial communities

3.1. Metagenomics

Next-generation sequencing technology has emerged over a decade ago, and has revolutionized the study of complex microbial communities, allowing for so called omics analyses (genomic, transcriptomic). Respectively, omics technologies are defined as technologies enabling the study of the DNA, RNA, proteins and metabolites from a specific environment. They can provide insights into microbial communities from diverse environments, such as the digestive tract of animals and humans (51–56), soil (57, 58), oceans (59, 60), as well as anaerobic reactors (61–63). More particularly, **metagenomics** (MG) has been widely used to explore the composition and the diversity of the human gut microbiome. Researchers have established that it is affected by various factors, such as age or geography (64).

Metagenomics analysis applied to the study of microbiomes is a cultivation-independent technology aiming at characterising the various microorganisms present in a specific environment. The principle of MG relies on sequencing the total DNA extracted from a given environment. Therefore, the total DNA from a sample is first extracted using an appropriate protocol (e.g. some of them include enrichment in bacterial and/or yeast DNA). Subsequently, the extracted DNA is cut into smaller fragments (between 300 and 800 pb) prior to sequencing. Sequencing results in the generation of millions of so called **metagenomic reads**. Usually, these reads are trimmed in order to remove those of low quality, and are later assembled into **contigs** (i.e *de novo* metagenomics reconstruction: larger DNA sequences assembled from sequencing reads based on the overlapping regions). Contigs are then analysed in order to predict the open reading frames (ORFs), using e.g. MetaProdigal (65), or other bioinformatic tools. Therefore, genes are searched on reconstructed contigs and are considered as unique. Finally, these ORFs can be functionally annotated by comparing with sequences registered in any available (public) database (e.g. GenBank (66), KEGG (67), COG (68), etc. …), or when submitted to an integrative database server for analysis and annotation (e.g. IMG-MER (69)).

Further processing of the re-constructed metagenomics contigs aims at generating **bins**, i.e. the grouping of contigs into one **metagenome assembled genome** (MAG) which will be further assigned to a taxonomic group. Binning can be performed using different available software, e.g. MetaBAT (70), Concoct (71) or MyCC (72), while taxonomic assignment is performed using e.g. PhyloPhlan (73).

When analyzing MG datasets, one common feature is the use of the metagenomic abundance. Metagenomic abundance refers to the abundance of a contig in a pool of sequenced DNA species. This is calculated by the number of reads that map to this contig and normalized by the contig length and per million mappable reads. Relative metagenomic abundance refers to the proportion (in percentage) of a specific contig in the studied sample (pool of contigs). Contigs originating from a common source (single genomes) should have similar average abundance, independently of their sizes. This feature is often used to bin contigs in a sequence-independent manner (so called sequence-independent binning) and is implemented in e.g. MetaBAT or CONCOCT softwares. It can be further used to calculate the average metagenomics abundance of re-constructed MAGs (relative to bacterial species) in the different samples.

Applied to various environments, MG allowed the characterization of microbial communities, as well as metabolic pathways, including the discovery of novel genes. Especially, MG allowed the identification and discovery of CAZymes from various environments, such as anaerobic reactors (74, 75), compost (76), termite's gut (77), crop of snails (78) as well as cow rumen (79). As an example, one of these studies includes the identification of new cellulases from a cellulose-degrading sludge of a lab-scale anaerobic reactor (80).

Additionally, MG studies applied to anaerobic reactors could help understanding how microorganisms adapt to the different environmental conditions applied to the reactors. For example, the microbial ecology of twelve mesophilic and thermophilic full-scale biogas plants treating manure or wastewater sludge was assessed and showed the identification of specific species related to different parameters such as feedstock and temperature (81). In line with other studies, the authors showed that the dominant phyla identified whitin the reactors were *Firmicutes*, *Bacteroidetes* and *Proteobacteria* (61, 63). However, MG generates huge datasets (including all genetic informations such as non-coding fragments), rendering the analysis of the genomic information complex and fastidious. On top of that, one limitation results from the

lack of knowledge on which function is really performed (i.e. is the gene expressed or just present?) by a microorganism in the environment under specific conditions. To this purpose, MG must be complemented with another approach to study the actual gene.

3.2. Metatranscriptomics

Metatranscriptomics (MT) is a technology aiming at studying the mRNA of an environmental sample, i.e. the expressed genes pool. Therefore, it can be seen as a technology complementary to MG as it can give additional informations regarding microbial pathways, and highly expressed genes (82, 83). Basically, the principle behind MT is quite similar to MG. Briefly, total mRNA from the studied environmental sample is extracted and sequenced, prior being assembled into mRNA contigs (gene transcripts), in the case of the *de novo* MT reconstruction. Subsequently, ORFs are predicted by a dedicated software (e.g. metaProdigal), followed by taxonomical and functional annotation of the resulting gene transcripts. Alternatively, if combined with a MG reconstruction, mRNA reads are mapped to the DNA template and relative expression/abundance of gene transcripts is calculated similarly to the DNA contigs abundance. This strategy is often referred to as **RNA-seq**.

Metatranscriptomics potentially reflects the effort of the organism and, when relevant, its associated microflora to break down LCB into useful compounds. For example, MT reports highlight the high representation and overexpression of cellulose and hemicelluloses degrading genes in the termite hindgut digestomes (84, 85). According to the MT analysis of hindgut paunch microbiota in wood- and dung-feeding termites, carbohydrate transport and metabolism were the second most expressed COG (Clusters of Orthologous Groups) function category, indicating the potential of MT to discover new CAZymes in this environment (86). Interestingly, the authors highlighted the value of the *de novo* metatranscriptome assembly in retrieving highly expressed genes (e.g. the top ten highly expressed glucose hydrolases were identified solely from the *de novo* assembled metatranscriptome and not from the accompanying metagenome), that are otherwise present at low relative abundance in metagenomes. The importance of *the de novo* metatranscriptome assembly, i.e. to recontruct and functionally characterize abundant transcripts, has been also discussed for other environments, including microbial communities in the deep-sea (87) and plankton

communities inhabiting surface and subpycnocline waters (88). Indeed, transcripts originating from the minority species were not well represented in the corresponding metagenomic datasets. Additionally, MT analysis was also applied to anaerobic reactors in order to highlight methanogenesis pathways (89, 90).

3.3. Metaproteomics and metabolomics

Metaproteomics and metabolomics correspond, respectively, to the technologies applied to analyse the proteins and the metabolites present in the studied environment. The typical workflow used in metaproteomics is the extraction and purification of the proteins from a sample followed by digestion into peptides before separation and mass spectrometry analysis (91). Then, the proteins present in the environment are identified by comparision of the experimental mass spectra to theorical mass spectra from protein databases. Various metaproteomic studies have been performed to characterizes the functional microbial community of diverse environments, such as anaerobic reactors (92, 93) as well as the human gut (94, 95). However, metaproteomics faces limitations related to the complexity of the datasets (extended computational time, software and hardware limitations), redundancy of the protein identification (identical peptides might belong to homologous proteins and can result in an ambiguous functional interpretation, (96) and finally, protein identification might be difficult due to the missing entries in protein databases (91). Finally, the study of metabolites, i.e. metabolomics, promises a huge potential for various areas (e.g. agronomy, medicine or environmental sciences). Yet, metabolomics studies are limited by many important issues. Indeed, there is no standardized extraction or analysis methods due to the diversity of metabolites having different chemical structures and properties, moreover metabolome is highly sensitive to environmental and genetic variations and finally, detection and data analysis remain tricky as the number of metabolites with a chemically known structrure is low (97). Therefore, in this PhD project, due to the limitations of metaproteomics and metabolomics mentioned above, I mainly relied on metagenomics and metatranscriptomics and their combination as tools to identify potentially interesting enzymes for the biorefinery sector.

4. Objectives and approach

The effective use of LCB for the biorefinery sector requires environmentally and energy efficient LCB pre-treatments. However, the current main thermal, chemical & mechanical pre-treatment processes are environmentally not friendly; alternative enzymatic pre-treatments are still inefficient and costly. Scientists have been putting efforts in finding new enzymes enabling a large-scale deconstruction of LCB.

In the context of bioprospecting for such enzymes, the aim of this PhD thesis was to characterize the carbohydrate hydrolytic potential of two microbiomes (restricted to bacteria). I focused on two different biological systems, i.e. anaerobic digestion and the termite's gut digestion system. Both systems are known to efficiently deconstruct LCB, thus being a source of CAZymes of interest. AD is a man-exploited process aiming at producing biogas from lignocellulosic biomass decomposition. The biomass to biogas conversion process results from the joint action of diverse microorganisms producing enzymes, including CAZymes. Next to AD, termites are insects living on lignocellulosic biomass. They developed a complex and efficient symbiosis with microorganisms to decompose lignocellulosic biomass.

Thus, harvesting the carbohydrate hydrolytic potential from these LCB degrading environments can help the discovery of novel enzymes and enzymes clusters with high potential for the biorefineries sector. Metagenomics and metatranscriptomics can be usefull tools to harvest this hydrolytic potential and identify the active enzymes. Further biochemical characterization of the new identified proteins is also important to determine their true hydrolytic activities.

In such an approach, several hypotheses have been proposed and tested within this PhD thesis (Figure 10):

- Combination of metagenomics and metatranscriptomics allows for the identification and characterization of key bacterial players involved in lignocellulose deconstruction
- Despite distinct lignocellulolytic communities, the bacterial communities from the termite gut and anaerobic digestion system show similarities at the level of the carbohydrate hydrolytic potential

- Accessory enzymes are also interesting to design nature-inspired cocktails dedicated to lignocellulose deconstruction; such cocktails are of interest to define deconstruction strategies to be exploited in the biorefinery sector

To test these hypotheses, the experimental approach undertaken was as follows (Figure 10):

- Optimize the protocols used for the heterologous production and biochemical characterization of enzymes identified in the following metagenomics and metatranscriptomics approaches (*discussed in annex 1)*
- Characterize the carbohydrate hydrolytic potential of an anaerobic reactor microbiome fed with lignocellulosic biomass (sugar beet pulp), using metagenomics and enzymatic studies of selected enzymes *(discussed in chapter 2)* and metagenomics combined with metatranscriptomics *(discussed in chapter 3)*
- Characterize the carbohydrate hydrolytic potential of termite's gut microbiome, fed with recalcitrant lignocellulosic biomass (miscanthus straw) using metagenomics combined with metatranscriptomics and followed by enzymatic studies of selected enzymes *(discussed in chapter 4)*
- Identify, heterologously produce and biochemically characterize novel carbohydrate active enzymes of potential interest for the biorefinery sector (*discussed in chapter 2, 3 and 4*)

CAZymes from lignocellulose deconstructing microbiomes

Hyp. 1	Combination of metagenomics and metatranscriptomics allows for the identification and characterization of key bacterial players involved in lignocellulose deconstruction
Hyp. 2	Despite distinct lignocellulolytic communities, the bacterial communities from the termite gut and anaerobic digestion system show similarities at the level of the carbohydrate hydrolytic potential
Hyp. 3	Accessory enzymes are also interesting to design nature-inspired cocktails dedicated to lignocellulose deconstruction; such cocktails are of interest to define deconstruction strategies to be exploited in the biorefinery sector

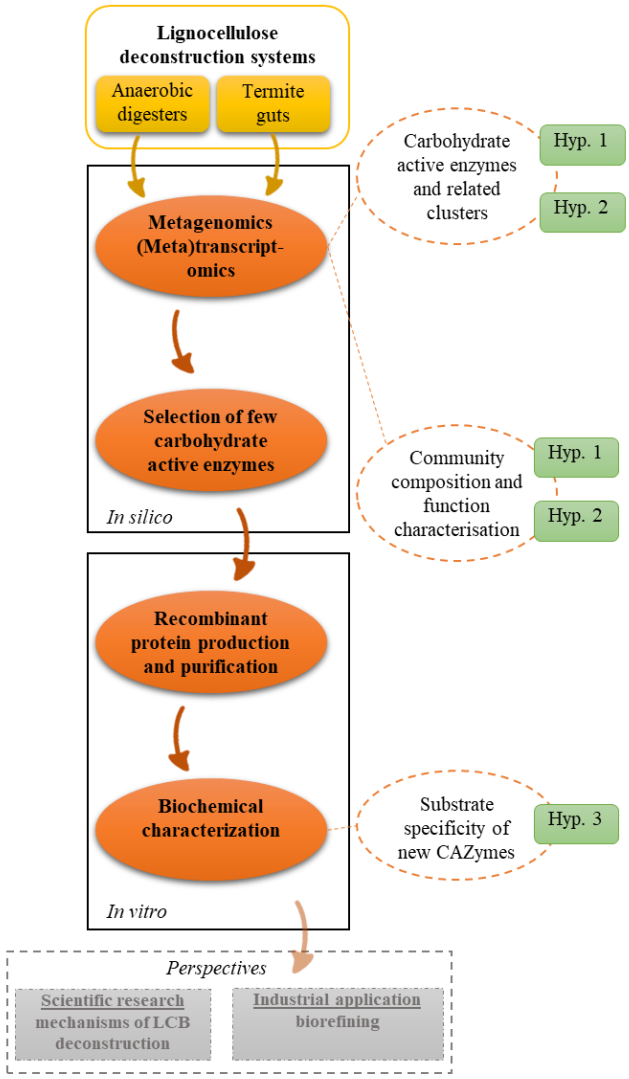


Figure 10: Schematic representation of the approach undertaken in the course of this study. Investigation of two different lignocellulose deconstruction systems by combining *in silico* analysis (metagenomics and metatranscriptomics) and *in vitro* analysis (recombinant protein production and biochemical characterization). The different hypotheses discussed in the frame of this PhD thesis are also represented. Hyp. refers to hypothesis

References

1. J. K. Saini, *et al.*, "Integrated Lignocellulosic Biorefinery for Sustainable Bio-Based Economy" in *Sustainable Approaches for Biofuels Production Technologies*, (Springer, Cham, 2019), pp. 121–146.
2. S. Malherbe, T. E. Cloete, Lignocellulose biodegradation: Fundamentals and applications. *Rev. Environ. Sci. Biotechnol.* **1**, 105–114 (2002).
3. B. S. Boneberg, *et al.*, Biorefinery of lignocellulosic biopolymers. *Rev. Eletrônica Científica da UERGS* **2**, 79 (2016).
4. S. Sun, S. Sun, X. Cao, R. Sun, The role of pretreatment in improving the enzymatic hydrolysis of lignocellulosic materials. *Bioresour. Technol.* **199**, 49–58 (2016).
5. J. Baruah, *et al.*, Recent trends in the pretreatment of lignocellulosic biomass for value-added products. *Front. Energy Res.* **6**, 1–19 (2018).
6. K. Jedvert, T. Heinze, Cellulose modification and shaping - A review. *J. Polym. Eng.* **37**, 845–860 (2017).
7. R. L. Whistler, "Hemicelluloses" in *Industrial Gums*, R. L. WHISTLER, J. N. BEMILLER, Eds. (1993), pp. 295–308.
8. T. Rezic, *et al.*, Integrated hydrolyzation and fermentation of sugar beet pulp to bioethanol. *J. Microbiol. Biotechnol.* **23**, 1244–1252 (2013).
9. M. Yadav, K. Paritosh, A. Chawade, N. Pareek, V. Vivekanand, Genetic engineering of energy crops to reduce recalcitrance and enhance biomass digestibility. *Agric.* **8** (2018).
10. Y. Zhang, *et al.*, Overcoming biomass recalcitrance by synergistic pretreatment of mechanical activation and metal salt for enhancing enzymatic conversion of lignocellulose. *Biotechnol. Biofuels* **12**, 1–15 (2019).
11. M. Carlsson, A. Lagerkvist, F. Morgan-Sagastume, The effects of substrate pre-treatment on anaerobic digestion systems: A review. *Waste Manag.* **32**, 1634–1650 (2012).
12. X. Zhang, C. Xu, H. Wang, Pretreatment of bamboo residues with Coriolus versicolor for enzymatic hydrolysis. *J. Biosci. Bioeng.* **104**, 149–151 (2007).
13. C. Wan, Y. Li, Microbial pretreatment of corn stover with Ceriporiopsis subvermispora for enzymatic hydrolysis and ethanol production. *Bioresour. Technol.* **101**, 6398–6403 (2010).
14. G. Banerjee, J. S. Scott-Craig, J. D. Walton, Improving enzymes for biomass conversion: A basic research perspective. *Bioenergy Res.* **3**, 82–92 (2010).
15. B. I. Cantarel, *et al.*, The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233–238 (2009).
16. M. D. Sweeney, F. Xu, Biomass Converting Enzymes as Industrial Biocatalysts for Fuels and Chemicals: Recent Developments. *Catalysts* **2**, 244–263 (2012).

17. R. Datta, *et al.*, Enzymatic degradation of lignin in soil: A review. *Sustain.* **9** (2017).
18. D. Guillén, S. Sánchez, R. Rodríguez-Sanoja, Carbohydrate-binding domains: Multiplicity of biological roles. *Appl. Microbiol. Biotechnol.* **85**, 1241–1249 (2010).
19. D. Floudas, *et al.*, The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science (80-.).* **336**, 1715–1719 (2012).
20. Z. Forsberg, *et al.*, Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing lytic polysaccharide monooxygenases. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8446–8451 (2014).
21. T. E. Takasuka, A. J. Book, G. R. Lewin, C. R. Currie, B. G. Fox, Aerobic deconstruction of cellulosic biomass by an insect-associated Streptomyces. *Sci. Rep.* **3**, 1–10 (2013).
22. W. T. Beeson, V. V. Vu, E. A. Span, C. M. Phillips, M. A. Marletta, Cellulose Degradation by Polysaccharide Monooxygenases. *Annu. Rev. Biochem.* **84**, 923–946 (2015).
23. J. Linares-Pasten, M. Andersson, E. Karlsson, Thermostable Glycoside Hydrolases in Biorefinery Technologies. *Curr. Biotechnol.* **3**, 26–44 (2014).
24. S. Comtet-Marre, *et al.*, Metatranscriptomics reveals the active bacterial and eukaryotic fibrolytic communities in the rumen of dairy cow fed a mixed diet. *Front. Microbiol.* **8** (2017).
25. B. Levy, E. Jami, Exploring the Prokaryotic Community Associated With the Rumen Ciliate Protozoa Population. *Front. Microbiol.* **9**, 1–14 (2018).
26. Z. Zhu, *et al.*, Changes in rumen bacterial and archaeal communities over the transition period in primiparous Holstein dairy cows. *J. Dairy Sci.* **101**, 9847–9862 (2018).
27. J. E. Graham, *et al.*, Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment. *Nat. Commun.* **2** (2011).
28. L. Artzi, E. A. Bayer, S. Moraïs, Cellulosomes: Bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15**, 83–95 (2017).
29. M. Garvey, H. Klose, R. Fischer, C. Lambertz, U. Commandeur, Cellulases for biomass degradation: comparing recombinant cellulase expression platforms. *Trends Biotechnol.* **31**, 581–93 (2013).
30. F. Thomas, J. H. Hehemann, E. Rebuffet, M. Czjzek, G. Michel, Environmental and gut Bacteroidetes: The food connection. *Front. Microbiol.* **2**, 1–16 (2011).
31. K. L. Anderson, A. A. Salyers, Biochemical Evidence that Starch Breakdown by Bacteroides thetaiotaomicron Involves Outer Membrane Starch-Binding Sites and Periplasmic Starch-Degrading Enzymes. **2**, 3192–3198 (1989).
32. M. K. Bjursell, E. C. Martens, J. I. Gordon, Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. *J. Biol. Chem.* **281**, 36269–36279 (2006).
33. E. C. Martens, N. M. Koropatkin, T. J. Smith, J. I. Gordon, Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. *J. Biol. Chem.* **284**, 24673–24677 (2009).
34. K. Tang, Y. Lin, Y. Han, N. Jiao, Characterization of potential polysaccharide utilization systems in the marine Bacteroidetes Gramella flava JLT2011 using a multi-omics approach. *Front. Microbiol.* **8**, 1–13 (2017).

35. V. Bagenholm, *et al.*, Galactomannan catabolism conferred by a polysaccharide utilization locus of Bacteroides ovatus: Enzyme synergy and crystal structure of a β-mannanase. *J. Biol. Chem.* **292**, 229–243 (2017).
36. J. M. Grondin, K. Tamura, G. Déjean, D. W. Abbott, H. Brumer, Polysaccharide Utilization Loci: Fuelling microbial communities. *J. Bacteriol.* **199**, JB.00860-16 (2017).
37. E. C. Martens, *et al.*, Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.* **9** (2011).
38. Á. T. Martínez, *et al.*, Biodegradation of lignocellulosics: Microbial, chemical, and enzymatic aspects of the fungal attack of lignin. *Int. Microbiol.* **8**, 195–204 (2005).
39. T. K. Lundell, M. R. Mäkelä, K. Hildén, Lignin-modifying enzymes in filamentous basidiomycetes - Ecological, functional and phylogenetic review. *J. Basic Microbiol.* **50**, 5–20 (2010).
40. A. Hatakka, Lignin-modifying enzymes from selected white-rot fungi: production and role from in lignin degradation. *FEMS Microbiol. Rev.* **13**, 125–135 (1994).
41. Y. Su, *et al.*, Biodegradation of lignin and nicotine with white rot fungi for the delignification and detoxification of tobacco stalk. *BMC Biotechnol.* **16**, 1–9 (2016).
42. R. Ten Have, P. J. M. Teunissen, Oxidative mechanisms involved in lignin degradation by white-rot fungi. *Chem. Rev.* **101**, 3397–3413 (2001).
43. T. D. H. Bugg, M. Ahmad, E. M. Hardiman, R. Rahmanpour, Pathways for degradation of lignin in bacteria and fungi. *Nat. Prod. Rep.* **28**, 1883–1896 (2011).
44. J. Larsbrink, *et al.*, A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
45. H. König, L. Li, J. Fröhlich, The cellulolytic system of the termite gut. *Appl. Microbiol. Biotechnol.* **97**, 7943–7962 (2013).
46. A. Brune, Symbiotic digestion of lignocellulose in termite guts. *Nat. Rev. Microbiol.* **12**, 168–180 (2014).
47. H. J. Guedegbe, *et al.*, Occurrence of fungi in combs of fungus-growing termites (Isoptera: Termitidae, Macrotermitinae). *Mycol. Res.* **113**, 1039–1045 (2009).
48. P. Talia, J. Arneodo, "Lignocellulose Degradation by Termites" in *Termites and Sustainable Management*, A. M. Khan, W. Ahmad, Eds. (2018), pp. 101–117.
49. A. Bayané, S. R. Guiot, Animal digestive strategies versus anaerobic digestion bioprocesses for biogas production from lignocellulosic biomass. *Rev. Environ. Sci. Biotechnol.* **10**, 43–62 (2011).
50. R. Ramaraj, N. Dussadee, Biological purification processes for biogas using algae cultures: A review. *Int. J. Sustain. Green Energy Int. J. Sustain. Green Energy. Spec. Issue Renew. Energy Appl. Agric. F. Nat. Resour. Technol.* **4**, 20–32 (2015).
51. R. Joynson, L. Pritchard, E. Osemwekha, N. Ferry, Metagenomic analysis of the gut microbiome of the common black slug arion ater in search of novel lignocellulose degrading enzymes. *Front. Microbiol.* **8**, 1–11 (2017).
52. N. Liu, *et al.*, Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *ISME J.*, 104–117 (2019).
53. W. Wang, H. Hu, R. T. Zijlstra, J. Zheng, M. G. Gänzle, Metagenomic

reconstructions of gut microbial metabolism in weanling pigs. *Microbiome* **7**, 1–11 (2019).
54. L. Wang, G. Zhang, H. Xu, H. Xin, Y. Zhang, Metagenomic analyses of microbial and carbohydrate-active enzymes in the rumen of holstein cows fed different forage-to-concentrate ratios. *Front. Microbiol.* **10** (2019).
55. J. Qin, *et al.*, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
56. M. Vital, *et al.*, Metagenomic insights into the degradation of resistant starch by human gut microbiota. *Appl. Environ. Microbiol.* **84**, 1–13 (2018).
57. V. Ahmed, M. K. Verma, S. Gupta, V. Mandhan, N. S. Chauhan, Metagenomic profiling of soil microbes to mine salt stress tolerance genes. *Front. Microbiol.* **9**, 1–11 (2018).
58. G. Feng, *et al.*, Metagenomic analysis of microbial community and function involved in cd-contaminated soil. *BMC Microbiol.* **18**, 1–13 (2018).
59. D. L. Kirchman, T. E. Hanson, M. T. Cottrell, L. J. Hamdan, Metagenomic analysis of organic matter degradation in methane-rich Arctic Ocean sediments. *Limnol. Oceanogr.* **59**, 548–559 (2014).
60. S. J. Biller, *et al.*, Data descriptor: Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 1–7 (2018).
61. J. Guo, *et al.*, Dissecting microbial community structure and methane-producing pathways of a full-scale anaerobic reactor digesting activated sludge from wastewater treatment by metagenomic sequencing. *Microb. Cell Fact.* **14**, 1–11 (2015).
62. Y. Yang, *et al.*, Metagenomic analysis of sludge from full-scale anaerobic digesters operated in municipal wastewater treatment plants. *Appl. Microbiol. Biotechnol.* **98**, 5709–5718 (2014).
63. X. Zhu, S. Campanaro, L. Treu, P. G. Kougias, I. Angelidaki, Novel ecological insights and functional roles during anaerobic digestion of saccharides unveiled by genome-centric metagenomics. *Water Res.* **151**, 271–279 (2019).
64. T. Bhattacharya, T. S. Ghosh, S. S. Mande, Global profiling of carbohydrate active enzymes in human gut microbiome. *PLoS One* **10**, 1–20 (2015).
65. D. Hyatt, P. F. Locascio, L. J. Hauser, E. C. Uberbacher, Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
66. D. A. Benson, *et al.*, GenBank. *Nucleic Acids Res.* **41**, 36–42 (2013).
67. M. Kanehisa, Y. Sato, K. Morishima, BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
68. M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Expanded Microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2015).
69. V. M. Markowitz, *et al.*, IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, 115–122 (2012).
70. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015**, 1–15 (2015).
71. J. Alneberg, *et al.*, Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
72. H. H. Lin, Y. C. Liao, Accurate binning of metagenomic contigs via automated

clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 12–19 (2016).
73. N. Segata, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes (2013) https:/doi.org/10.1038/ncomms3304.PhyloPhlAn.
74. Y. Wei, *et al.*, Insight into dominant cellulolytic bacteria from two biogas digesters and their glycoside hydrolase genes. *PLoS One* **10**, 1–19 (2015).
75. S. Wongwilaiwalin, *et al.*, Comparative metagenomic analysis of microcosm structures and lignocellulolytic enzyme systems of symbiotic biomass-degrading consortia. *Appl. Microbiol. Biotechnol.* **97**, 8941–8954 (2013).
76. M. J. Dougherty, *et al.*, Glycoside Hydrolases from a targeted Compost Metagenome, activity-screening and functional characterization. *BMC Biotechnol.* **12** (2012).
77. G. Bastien, *et al.*, Mining for hemicellulases in the fungus-growing termite Pseudacanthotermes militaris using functional metagenomics. *Biotechnol. Biofuels* **6**, 1–15 (2013).
78. A. M. Cardoso, *et al.*, Metagenomic Analysis of the Microbiota from the Crop of an Invasive Snail Reveals a Rich Reservoir of Novel Genes. *PLoS One* **7** (2012).
79. M. Hess, *et al.*, Metagenomic Discovery of Biomass-Degrading Genes and genomes from Cow Rumen. *Science (80-.).* **463**, 463–467 (2011).
80. Y. Xia, F. Ju, H. H. P. Fang, T. Zhang, Mining of Novel Thermo-Stable Cellulolytic Genes from a Thermophilic Cellulose-Degrading Consortium by Metagenomics. *PLoS One* **8** (2013).
81. S. Campanaro, L. Treu, P. G. Kougias, G. Luo, I. Angelidaki, Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water Res.* **140**, 123–134 (2018).
82. J. A. Gilbert, *et al.*, Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *PLoS One* **3**, 1–13 (2008).
83. T. Urich, *et al.*, Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3** (2008).
84. A. Tartar, *et al.*, Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite Reticulitermes flavipes. *Biotechnol. Biofuels* **2**, 1–19 (2009).
85. R. Raychoudhury, *et al.*, Comparative metatranscriptomic signatures of wood and paper feeding in the gut of the termite Reticulitermes flavipes (Isoptera: Rhinotermitidae). *Insect Mol. Biol.* **22**, 155–171 (2013).
86. S. He, *et al.*, Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. *PLoS One* **8** (2013).
87. B. J. Baker, *et al.*, Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *ISME J.* **7**, 1962–1973 (2013).
88. I. Hewson, *et al.*, Metatranscriptomic analyses of plankton communities inhabiting surface and subpycnocline waters of the chesapeake bay during oxic-anoxic-oxic transitions. *Appl. Environ. Microbiol.* **80**, 328–338 (2014).
89. V. Nolla-Ardevol, M. Strous, H. E. Tegetmeyer, Anaerobic digestion of the microalga Spirulina at extreme alkaline conditions: Biogas production,

metagenome and metatranscriptome. *Front. Microbiol.* **6** (2015).
90. Y. Xia, *et al.*, Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis. *Sci. Rep.* **4**, 1–9 (2014).
91. R. Heyer, *et al.*, Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36 (2017).
92. R. Heyer, *et al.*, Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. *Biotechnol. Biofuels* **9**, 1–16 (2016).
93. F. Abram, *et al.*, A metaproteomic approach gives functional insights into anaerobic digestion. *J. Appl. Microbiol.* **110**, 1550–1560 (2011).
94. C. A. Kolmeder, *et al.*, Faecal metaproteomic analysis reveals a personalized and stable functional microbiome and limited effects of a probiotic intervention in adults. *PLoS One* **11**, 1–23 (2016).
95. C. A. Kolmeder, *et al.*, Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS One* **7** (2012).
96. F. A. Herbst, *et al.*, Enhancing metaproteomics-The value of models and defined environmental microbial systems. *Proteomics* **16**, 783–798 (2016).
97. F. Courant, J.-P. Antignac, G. Dervilly-Pinel, B. Le Bizec, Basics of mass spectrometry based metabolomics. *Proteomics* **14**, 2369–2388 (2014).
98. A. Berlin, No barriers to cellulose breakdown. *Science (80-.).* **342**, 1454–1456 (2013).
99. O. Yaniv, *et al.*, Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of Clostridium thermocellum anti-σI factors. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 522–534 (2014).

Chapter 2

Carbohydrate Hydrolytic Potential and Redundancy of an Anaerobic Digestion Microbiome Exposed to Acidosis, as revealed by Metagenomics

Presented as a slightly modified version of the journal article published as:
Bertucci M., Calusinska M., Goux X., Rouland-Lefèvre C., Untereiner B., Ferrer P., Gerin P. A., Delfosse P. (2019). Carbohydrate Hydrolytic Potential and Redundancy of an Anaerobic Digestion Microbiome Exposed to Acidosis, as Uncovered by Metagenomics. *Applied and Environmental Microbiology*, 85:e00895-19.

Personal contribution:

- Characterization of the carbohydrate hydrolytic potential using metagenomic dataset generated by another co-author.
- Experimental design and lab-work (recombinant protein production, purification, enzymatic assays, etc.), with the help of co-authors.
- Redaction of the full chapter including the comments and corrections of all the co-authors

In Chapter 2, metagenomics has been applied to a microbiome from an anaerobic digestion reactor excessively fed with sugar beet pulp (a substrate selected for its high polysaccharide content) in order to reach acidosis (pH drop from 7 to 5.5). The aim was to determine the bacterial community composition during the whole process as well as its carbohydrate hydrolytic potential. For this work, I relied on high-throughput DNA sequencing and downstream bioinformatic analyses. As a result, I could highlight, besides the changing community composition, the functional redundancy of this bacterial community for the hydrolysis of carbohydrates (retention of the functional potential despite the change in community composition) along the full course of the experiment (266 days) and even at low pH (reactor under acidosis). To complement this study, I further selected from the reconstructed metagenome several genes coding for carbohydrate active enzyme, co-localised in a gene cluster (*in silico* identified as targeting glucomannan). I also succeeded in the heterologous production of the corresponding proteins and assessed their activity. The main outcomes of the work reported in this chapter are 1) an insight in anaerobic bacterial strategies to deconstruct lignocellulose biomass, 2) a model for the mode of action of an acetylated glucomannan-targeting PUL, and 3) the suggestion of including accessory enzymes in cocktails to improve the deconstruction of lignocellulose in the biorefinery sector.

Abstract

Too fast hydrolysis and fermentation of easily digestible biomass substrates may lead to acidosis of anaerobic reactors and decreased methane production. Previously, it was shown that the structure of microbial communities changes during acidosis; however, once the conditions are back to optimal, biogas (initially CO_2) production quickly restarts. This suggests the retention of the community functional redundancy during the process failure. In this study, with the use of metagenomics and downstream bioinformatics analyses, we characterize the carbohydrate hydrolytic potential of the bacterial community, with a special focus on the evolution during acidosis. To that purpose, genes encoding for carbohydrate-active enzymes were identified, and to further link the community hydrolytic potential with key microbes, bacterial genomes were reconstructed. In addition, we produced and characterized biochemically the specificity and activity of selected enzymes, thus verifying the accuracy of the *in silico* predictions. The results confirm the retention of the community hydrolytic potential during acidosis and indicate *Bacteroidetes sp.* to be largely involved in biomass degradation. *Bacteroidetes* showed high diversity and genomic content of carbohydrate hydrolytic enzymes that might favor the dominance of this phylum over other bacteria in some anaerobic reactors. The combination of bioinformatic analyses and activity tests enabled us to propose a model of acetylated glucomannan degradation by *Bacteroidetes*.

Importance

The enzymatic hydrolysis of lignocellulosic biomass is mainly driven by the action of carbohydrate-active enzymes. In this study, by characterizing the gene profiles at the different stages of the anaerobic digestion experiment, we showed that the microbiome retains its hydrolytic functional redundancy even during severe acidosis, despite significant changes in taxonomic composition. By analyzing reconstructed bacterial genomes, we observe a large diversity in *Bacteroidetes* hydrolytic genes that likely supports the abundance of this phylum in some anaerobic digestion systems. Further, we observe genetic redundancy within the *Bacteroidetes* group, which accounts for the preservation of the hydrolytic potential during acidosis. This work also reveals new polysaccharide utilization loci involved in the deconstruction of

various biomasses and proposes a model of acetylated glucomannan degradation by *Bacteroidetes*. Acetylated glucomannan-enriched biomass is a common substrate for many industries, including pulp and paper production. Using naturally evolved cocktails of enzymes for biomass pretreatment could be an interesting alternative to the commonly used chemical pretreatments.

Keywords: *Bacteroidetes*, biotechnology, enzymes, molecular biology, polysaccharides, recombinant-protein production

1. Introduction

Anaerobic digestion (AD) of biomass (including biowaste) is a process aiming at producing biogas, i.e. a mixture of methane (CH_4), carbon dioxide CO_2), and trace gases (i.e., H_2, NH_3, and H_2S) (1). Biogas is a result of the joint action of diverse microorganisms that act synergistically to deconstruct organic matter. In the first stage, namely hydrolysis, the organic matter (i.e., fats, proteins, and polysaccharides) is deconstruct by fermentative bacteria into soluble molecules. During the acidogenesis and acetogenesis stages, bacterial consortia convert the resulting monomers and oligomers mainly into CO_2, H_2 and volatile fatty acids (VFAs), including acetate. Finally, methane is produced by methanogenic archaea during the methanogenesis stage (2). Even though different environmental factors and operational conditions can lead to the dysfunctioning of anaerobic digesters, acidification caused by an accumulation of VFAs (here, referred to as acidosis) is one of the most often recorded phenomena (3, 4). It results from the decoupling of the hydrolytic and acidogenic stages (which perform too fast due to, e.g., higher multiplication rates of involved organisms under specific conditions) from the downstream acetogenesis and methanogenesis stages (which are too slow due to, e.g., slower multiplication rates of the acetogenic bacteria and methanogenic archaea) (5). As reported from previous studies, acidosis leads to a change in the microbial community structure, i.e., the taxonomic composition (4, 6). Both bacterial and archaeal communities are affected, and, as a result, methane production is slowed down and sometimes interrupted. Interestingly, following a recovery stage, the newly established community relatively quickly restarts the biogas production (3, 7). This suggests a functional community redundancy (similar metabolic functions performed by distinct coexisting microorganisms) (8) of the AD microbial community during the acidosis.

Hydrolysis is often considered a bottleneck of the AD process because it either underperforms (in the case of the recalcitrant biomass, e.g., highly lignified or high crystalline cellulose content) or outperforms (fast hydrolysis of an easily digestible biomass can lead to acidosis). The enzymatic hydrolysis of lignocellulose biomass is mainly driven by the action of carbohydrate active enzymes (CAZymes) and especially by glycoside hydrolases (GHs), carbohydrate esterases (CEs),

polysaccharide lyases (PLs), and other auxiliary enzymes (AAs) (9). Carbohydrate-binding modules (CBMs) are noncatalytic contiguous amino acid sequences, which bind hydrolytic enzymes to their carbohydrate substrates. They usually exist as modules within larger enzymes, but some can be independent proteins (10). Glycosyltransferases (GTs) are also classified within the CAZy database as they are involved in the biosynthesis of glycosidic bonds from phosphate-activated sugars donors. GHs hydrolyze and/or transglycosylate the glycosidic bonds, CEs hydrolyze ester bonds, and AAs are redox enzymes acting in conjunction with other CAZymes, many of which are involved in lignin degradation. PLs cleave bonds from uronic acid-containing polysaccharide chains (nonhydrolytic cleavage of glycosidic bonds). With the emergence of the high-throughput sequencing technologies, owing to the reduced cost and easier access, researchers have investigated the metabolic potential (referring to the gene content) of the microbial communities present in anaerobic environments. In a variety of studies, microbial communities of full-scale and lab-scale anaerobic digesters under optimal (11) and dysfunctioning conditions have been characterized (4, 12–14). Among bacteria, *Bacteroidetes* and *Firmicutes* have often been cited as the most dominant phyla, and their genomes were shown to contain high number of CAZyme-coding genes (15–17). CAZyme-coding genes can be expressed separately in bacterial genomes or they can form operons of coexpressed genes, e.g., cellulosomes in *Firmicutes* or polysaccharide-utilization loci (PULs) in *Bacteroidetes*. While cellulosomes have been given large scientific attention in the past (18, 19), PULs were much less investigated (18, 20). PULs are defined as gene clusters encoding cell envelope-associated enzymes that allow *Bacteroidetes* to bind to and degrade a specific polysaccharide and to import the released oligosaccharides inside the cell (21, 22). They are composed of at least one pair of susC and susD genes encoding a TonB-dependent transporter (TBDT) and a cell-surface glycan-binding protein (SGBP), respectively (23). The protein SGBP is involved in the binding of carbohydrates while TBDT transports them through the outer membrane. The first PUL was identified from *Bacteroides thetaiotaomicron* isolated from the human gut and is dedicated to starch degradation (24). During the last decades, a number of PULs have been discovered from omics studies, unravelling a wider diversity of CAZyme-coding genes than previously thought (25–29). One study from the marine environment, reported the characterization of three PULs (two targeting xylan and one

targeting pectins) using combined omics approaches, i.e., transcriptomics, proteomics and metabolomics (30). Another study identified a xyloglucan- specific PUL (from the human gut) through a metagenomics approach that was further experimentally validated (31).

The general purpose of the present study was to characterize the hydrolytic potential of the bacterial community in an anaerobic digestion reactor. In particular, we aimed at assessing the carbohydrate hydrolytic potential in the context of an acidosis event, which is a common type of reactor dysfunction. Previous studies have shown that acidosis impacts microbial community structure at the taxonomic level. Here, we wanted to assess the community functional redundancy, with a special focus on the carbohydrate hydrolytic capacity of the anaerobic digester community. Metagenomic reconstruction allowed identification of 4,148 genes putatively assigned to CAZy families (including 49 putative PULs); only 1,052 of these were further assigned an Enzyme Commission (EC) class related to a potential hydrolytic function. To further link the community hydrolytic potential with key microbes, we investigated the newly reconstructed metagenome-assembled genomes (MAGs). Additional enzymatic activity assays were performed on a subset of predicted CAZymes belonging to a PUL named PUL219 from a dominant *Bacteroidetes* MAG in order to confirm their predicted activities.

2. Material and methods

2.1. Sampling, metagenomics, and data processing

Samples were taken from an R3 lab-scale continuously stirred tank reactor (CSTR) of 100-liter capacity fed with dry sugar beet pulp as described in a previous study (4). Decreasing HRT and increasing OLR resulted in a pH decrease and lower biomass-to -methane conversion ratio (4). In total, seven time points were analyzed, corresponding to the different stages of the experiment (Figure 1). Total genomic DNA was extracted with a PowerSoil DNA isolation kit (MoBio Laboratories, Inc.), according to the manufacturer's instructions. Metagenomic DNA libraries were constructed, sequenced, and analyzed as previously described (32). Binning of assembled contigs resulted in over 30 MAGs with average genome completeness of 77% ± 20% and contamination below 5%. MAGs described in this study correspond

to the bacterial bins listed in Table 2 in a previous publication (32). The taxonomic assignment of MAGs was done with Phylophlan (33). Prodigal was used to predict coding sequences on the coassembled contigs (34). CAZyme-coding genes were searched with the dbCAN (dbCAN-fam-HMMs.txt.v6 (35)) against a CAZy database (http://www.cazy.org/). The resulting CAZyme-coding genes were manually curated. A difference in abundances of MAGs/GHs/ECs was tested by a chi-square test. PULs were predicted according to the PUL database (36), and the presence of at least one susD or susC gene was mandatory for partial PUL prediction. Homology to peptide pattern (Hotpep) was used to assign the identified CAZymes to an EC class (37). Taxonomic and functional annotation of the coassembled contigs was done with IMG-MER (38) and corresponds to the project identification number 3300002079. The most complete MAGs were reannotated with RAST (39).

2.2. CAZyme-coding gene heterologous expression in *E. coli*

Details for gene selection, isolation, and cloning are shown in Text S1 in the supplementary material. For the expression of CAZyme-coding gene, the plasmid pET-52b(+) (Millipore Corporation, Billerica, MA, USA) and the E. *coli* Rosetta(DE3) strain (Millipore Corporation, Billerica, MA, USA) were used. Genes were ligated into pET-52b(+) vector and introduced in E. *coli* by heat shock. An isolated colony (the insert was determined by sequencing) was grown overnight at 37°C in LB medium containing ampicillin (LB-Amp) shaking at 250 rpm. In total, 100 µl of this culture was transferred into a fresh 100 ml of LB-Amp. Cells were grown at 37°C with shaking at 250 rpm to an optical density at 600 nm (OD600) of 0.5, and isopropyl-β-D-thiogalactopyranoside (IPTG; ThermoFisher, Waltham, MA, USA) was added at final concentration of 0.5 mM to induce the expression of recombinant proteins. The culture was maintained for 20 h at room temperature and 250 rpm.

Cells were collected by centrifugation at 5,000 × g at 4°C for 15 min. Supernatant was stored at -20°C, and the cell pellet was resuspended in an appropriate lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8). As a negative control, an empty pET-52b(+) vector was cloned, expressed, and processed as a recombinant protein sample.

Using a method adapted from previous studies (40, 41), cell lysis was achieved using a Sonicator VC750 (Sonics & Materials, Inc., Newtown, CT, USA) with a 2-min pulse (1 s on/1s off) followed by a 2-min pause (40% of amplitude). This step of sonication was repeated twice. Samples were transferred into 2-ml tubes and centrifuged for 15 min at 16,000 × g at 4°C. The liquid part was filtered using an Acrodisc 13-mm syringe filter with a 0.2-µm-pore-size Supor membrane (Millipore Corporation, Billerica, MA, USA). Protein detection and partial purification were performed using the histidine tag located at the C terminus. Initial detection in cell lysates was performed using Western blotting. Separation was done using 10% Mini-Protean TGX precast protein gels (Bio-Rad, Hercules, CA, USA). Transfer was performed with a Trans-Blot Turbo transfer system (Bio-Rad, Hercules, CA, USA). Detection was carried out using an appropriate antibody labeled with horseradish peroxidase (HRP) (6×His tag polyclonal antibody) (ThermoFisher, Waltham, MA, USA). nickel-nitrilotriacetic acid (NTA) agarose matrix was used to load polypropylene columns for partial purification according to the supplier's instructions (Qiagen, Hilden, Germany). Protein quantification was determined using a reducing agent and detergent-compatible (RC DC) protein assay (Bio-Rad, Hercules, USA) and bovine serum albumin as a standard. Proteins were identified, and details can be found in Text S2 in the supplementary material.

2.3. Signal peptide prediction and activity assays

Signal peptides were predicted using LipoP, version 1.0 (http://www.cbs.dtu.dk/services/LipoP/) (42). Enzymatic assays were performed using 4-nitrophenyl derivatives as substrates (Sigma-Aldrich, Saint Louis, MO, USA). Initially, 50 µl of partially purified protein solution was incubated with 100 µl of citrate phosphate buffer, pH 7 (0.1 M citric acid, 0.2 M dibasic sodium phosphate), and 50 µl of substrate (Table S1 in the supplementary material for substrate concentration). A microplate was incubated at 37°C in a Tecan Spark 20 M microplate reader (Tecan, Mannedorf, Switzerland). Th rate of released 4-nitrophenol was monitored at 405 nm. Assays were performed in triplicate.

Following previous studies (43, 44), enzymatic activity was also assessed by the release of reducing sugars. CMC (Sigma-Aldrich, Saint Louis, USA), arabinoxylan, galactomannan, konjac glucomannan, and xylan (Megazyme, Wicklow, Ireland) were

used as substrates. Briefly, 100 µl of partially purified enzyme solution was incubated with 50 µl of citrate phosphate buffer, pH 7 (0.1 M citric acid, 0.2 M dibasic sodium phosphate), and 100 µl of substrate at 37°C for 30 min (Table S1 in the supplementary material for substrate concentration). The concentrations of reducing sugars were determined applying the Somogyi-Nelson method, and absorbance was read at 620 nm using Specord Plus spectrophotometer (Analytik Jena, Jena, Germany). Assays were performed in triplicates.

Single enzymes as well as enzymatic cocktails were used for the pretreatment of glucomannan to determine the release of D-glucose, D-mannose, and acetic acid. Briefly, pretreatment was carried out in 500 µl as follows: 250 µl of substrate (Table S1 in the supplementary material for substrate concentration), 50 µl of buffer, and 50 µl of each enzyme assessed; final volume was reached by addition of water if necessary. The reaction mixture was incubated at 37°C for 1 h. If a cascade reaction was tested, a second enzyme (enzymatic set) was added after 1 h of pretreatment and reincubated for 1 h at 37°C. The release of D-glucose, D-mannose, and acetic acid was determined using commercially available kits (Megazyme, Wicklow, Ireland), according to the supplier's instructions. An enzymatic cocktail containing uBac-GH3, uBac-CE6, uBac-GH26a, and uBac-GH26b was taken as a reference to normalize results to 100%.

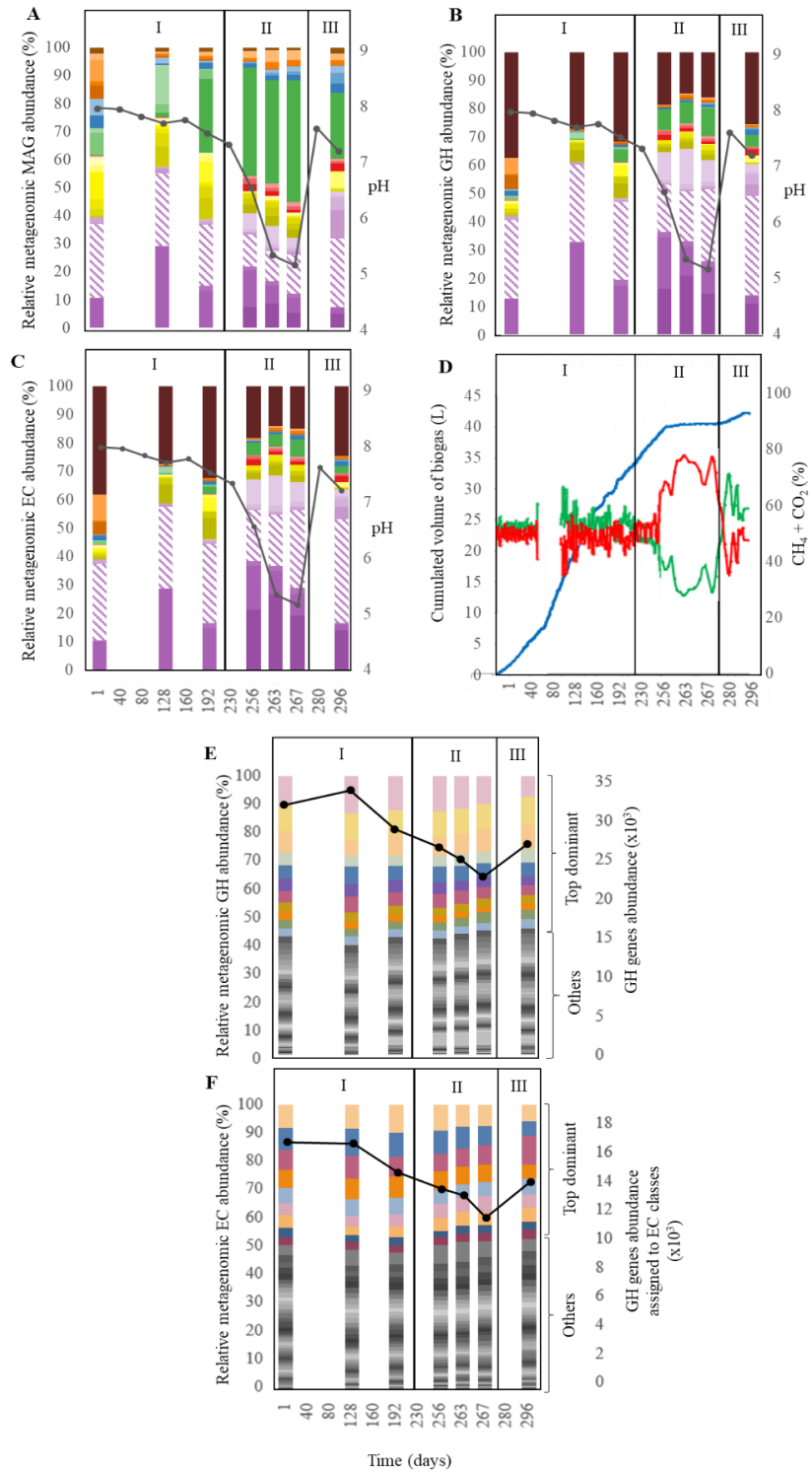
2.4. Data availability

The most complete MAGs can be accessed using a RAST (http://rast.nmpdr.org) guest account under the following identification numbers: 6666666.364478 (MAG1), 6666666.364479 (MAG2), 6666666.364480 (MAG3), 6666666.364481 (MAG4), 6666666.364482 (MAG5), 6666666.364483 (MAG6), 6666666.364484 (MAG7), 6666666.364485 (MAG8), 6666666.364486 (MAG9), 6666666.364487 (MAG10), 6666666.364489 (MAG11), 6666666.364490 (MAG12), 6666666.364491 (MAG13), 6666666.364493 (MAG14), 6666666.364494 (MAG15), 6666666.364495 (MAG16), 6666666.364496 (MAG17), 6666666.364497 (MAG18), 6666666.364498 (MAG19), 6666666.364499 (MAG20), 6666666.364500 (MAG21), 6666666.364502 (MAG23), 6666666.364503 (MAG25), 6666666.364504 (MAG30), and 6666666.364505 (MAG31).

3. Results

3.1. Functional redundancy of hydrolytic metabolism under changing environmental conditions in AD reactors

In our previous study (4), the dynamic of the microbial composition of the community was investigated with the use of high-throughput 16S rRNA gene amplicon sequencing in triplicate lab-scale AD reactors operated for 300 days. Reactors were sequentially exposed to a decreasing hydraulic retention time (HRT) and increasing organic loading rate (OLR; phase I) of dry sugar beet pulp, leading to acidosis (phase II) and process recovery (phase III), as indicated in Figure 1. Here, the hydrolytic potential of the bacterial community from one of the reactors (named R3 in (4)) was further investigated with the use of metagenomics. Genomic DNA was extracted from seven samples taken at different stages of the experiment and sequenced using a high-throughput approach (32). As a result, over 30,000 contigs were reconstructed. Out of the 75,002 nonredundant protein coding genes identified in the reconstructed contigs, 4,148 (representing 5.53% of all coding sequences) were assigned to CAZy families (dbCAN analysis) (35), representing GH, GT, AA, CE, and PL, as well as CBM classes. In total, 1,052 putative CAZymes were further functionally classified into EC protein categories (Hotpep analysis) (37). At the microbiome level, we found a good correlation between the average metagenomic GH coding gene abundance at the different stages of the experiment and the absolute number of genes assigned to the corresponding family ($R2 = 0.89$, $P < 0.001$) (Figure 2A).


A
B
C
D
E
F
I
II
III
Relative metagenomic MAG abundance (%)
Relative metagenomic GH abundance (%)
Relative metagenomic EC abundance (%)
pH
Cumulated volume of biogas (L)
$CH_4 + CO_2$ (%)
Top dominant
Others
GH genes abundance ($x10^3$)
GH genes abundance assigned to EC classes ($x10^3$)
Time (days)

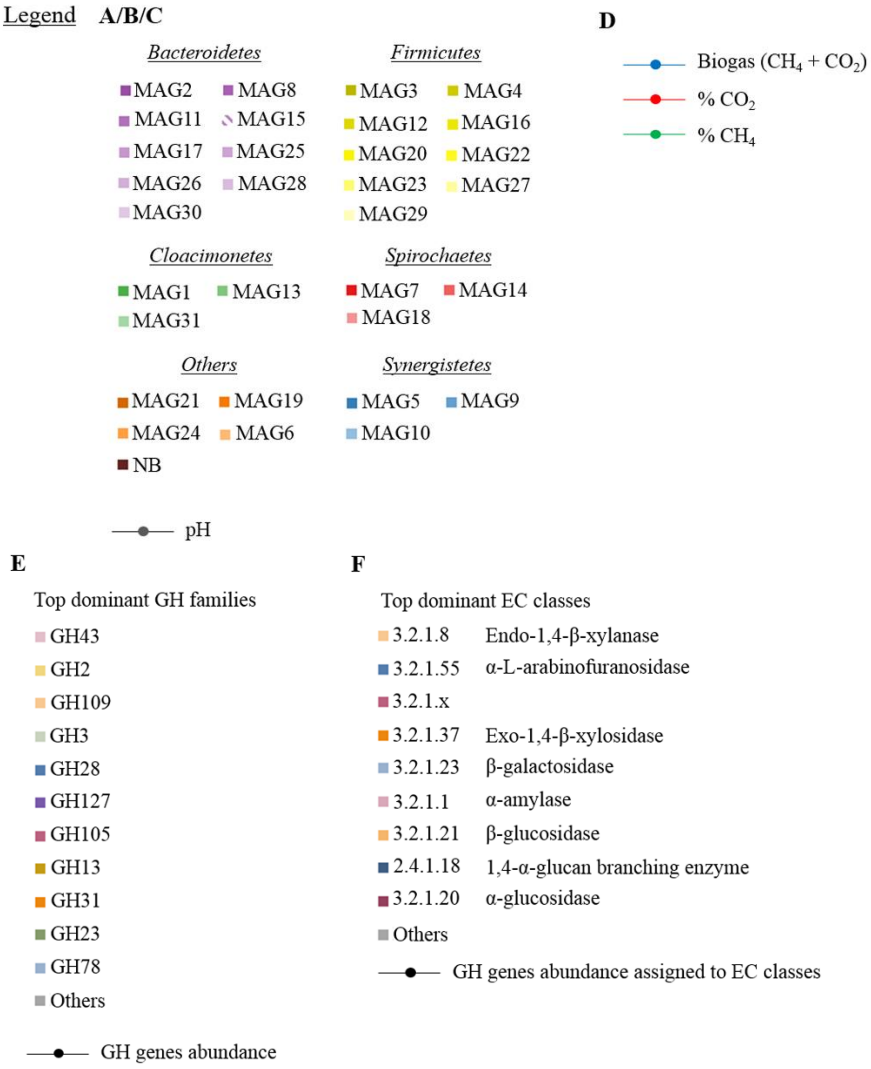

Legend A/B/C
Bacteroidetes
MAG2 MAG8 MAG11 MAG15 MAG17 MAG25 MAG26 MAG28 MAG30
Firmicutes
MAG3 MAG4 MAG12 MAG16 MAG20 MAG22 MAG23 MAG27 MAG29
Cloacimonetes
MAG1 MAG13 MAG31
Spirochaetes
MAG7 MAG14 MAG18
Others
MAG21 MAG19 MAG24 MAG6 NB
Synergistetes
MAG5 MAG9 MAG10
pH
D
Biogas (CH4 + CO2)
% CO2
% CH4
E
Top dominant GH families
GH43
GH2
GH109
GH3
GH28
GH127
GH105
GH13
GH31
GH23
GH78
Others
GH genes abundance
F
Top dominant EC classes
3.2.1.8 Endo-1,4-β-xylanase
3.2.1.55 α-L-arabinofuranosidase
3.2.1.x
3.2.1.37 Exo-1,4-β-xylosidase
3.2.1.23 β-galactosidase
3.2.1.1 α-amylase
3.2.1.21 β-glucosidase
2.4.1.18 1,4-α-glucan branching enzyme
3.2.1.20 α-glucosidase
Others
GH genes abundance assigned to EC classes

Figure 1: Metagenomics-assisted characterization of bacteria and their hydrolytic potential throughout the anaerobic digestion (AD) experiment. **A** - Relative metagenomic abundance of the different metagenome-assembled genomes (MAGs). **B** - Relative metagenomic abundance of the different glycoside hydrolases (GH) colored according to the reconstructed MAG. **C** - Relative metagenomic abundance of the different GHs additionally assigned to an enzyme commission (EC) category, colored according to the reconstructed MAG. **D** - Characteristics of the biogas produced, including the cumulative volume of biogas produced and its composition in CH_4 (%) and CO_2 (%) at the different stages of the experiment. **E** - Relative metagenomic abundance of the different GHs colored according to the GH family. The top dominant families further discussed in the manuscript are highlighted. The GH gene abundance, at the microbiome (whole-community) level, is represented by a black solid line. **F -** Relative metagenomic abundance of the different GHs additionally assigned to an EC category, colored according to the EC category. The top dominant categories further discussed in the manuscript are highlighted. GH genes abundance assigned to EC classes is represented by a black solid line. Roman numerals indicate phases I to III of the AD experiment, as follows: I, decreasing hydraulic retention time (HRT) and increasing organic loading rate (OLR) of sugar beet pulp; II, acidosis; III, process recovery. NB, not binned.

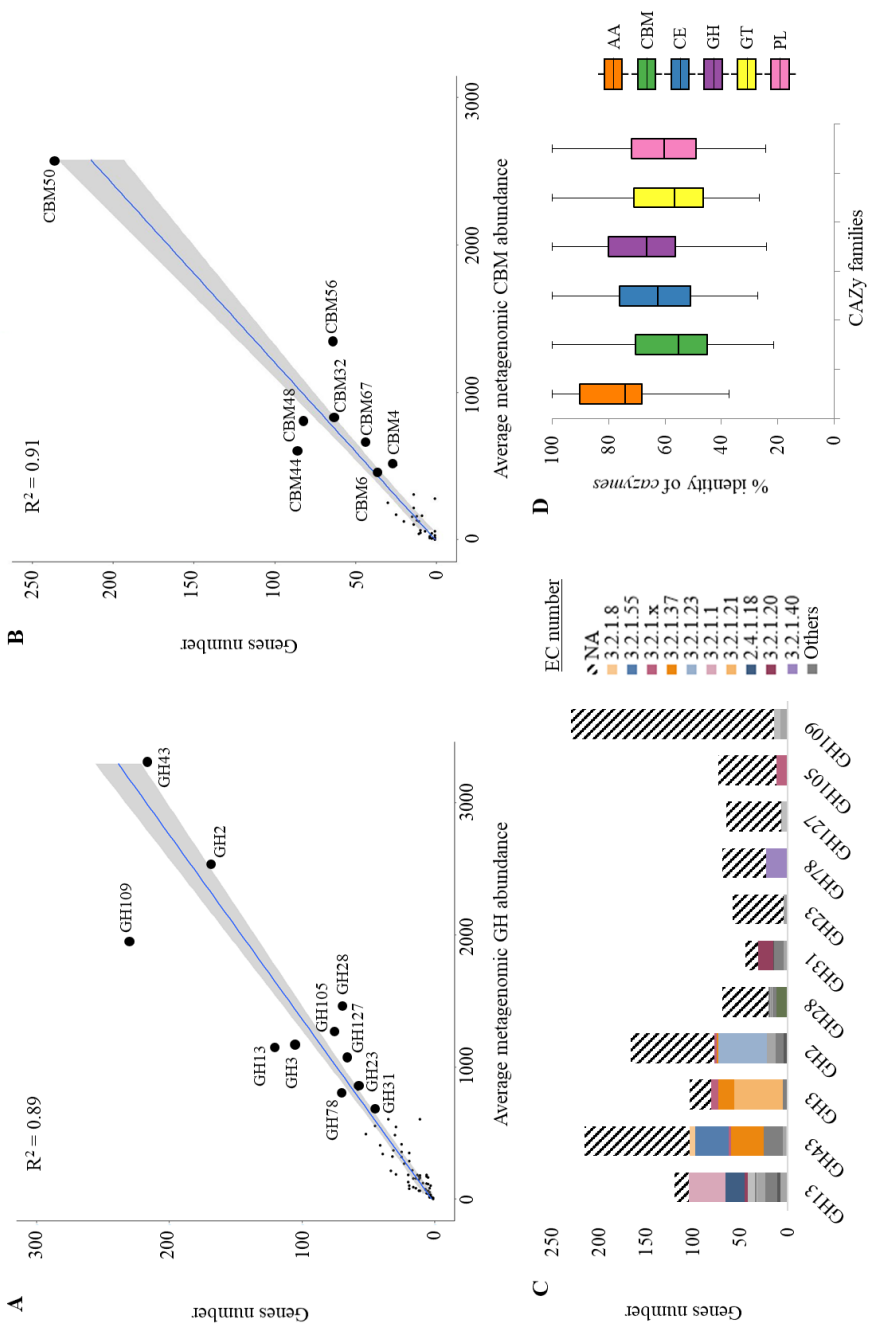

A
Genes number
Average metagenomic GH abundance
R² = 0.89
GH109
GH43
GH2
GH13
GH3
GH105
GH28
GH127
GH23
GH78
GH31
B
Genes number
Average metagenomic CBM abundance
R² = 0.91
CBM50
CBM56
CBM48
CBM32
CBM44
CBM67
CBM6
CBM4
C
Genes number
EC number
NA
3.2.1.8
3.2.1.55
3.2.1.x
3.2.1.37
3.2.1.23
3.2.1.1
3.2.1.21
2.4.1.18
3.2.1.20
3.2.1.40
Others
GH13
GH43
GH3
GH2
GH28
GH31
GH23
GH127
GH105
GH78
GH109
D
% identity of *cazymes*
CAZy families
AA
CBM
CE
GH
GT
PL

Figure 2: Characteristics of CAZy-coding genes identified in the anaerobic digestion (AD) reactor. **A** - Correlation between the average metagenomic glycoside hydrolase (GH) abundance and the absolute number of genes assigned to the corresponding family at the microbiome (whole-community) level. Top dominant families further discussed in the manuscript are highlighted. **B** - Correlation between the average metagenomic carbohydrate binding module (CBM) abundance and the absolute number of gene assigned to the corresponding family at the microbiome level. Top dominant CBM families are highlighted. **C** - Assignment of the top dominant GHs to an enzyme commission (EC) category. NA, not assigned; EC 3.2.1.8, endo-1,4-þ-xylanase; EC 3.2.1.55, a-L-arabinofuranosidase; EC 3.2.1.37, þ-xylosidase; EC 3.2.1.23, þ-galactosidase; EC 3.2.1.1, a-amylase; EC 3.2.1.21, þ-glucosidase; EC 2.4.1.18, 1,4-a-glucan branching enzyme; EC 3.2.1.20, a-glucosidase; EC 3.2.1.40, a-L-rhamnosidase. **D** - Box plot representations of the percentage identity of the CAZymes to proteins in the NCBI nonredundant protein database. The different CAZy families represented include auxiliary enzymes (AA), carbohydrate-binding modules (CBM), carbohydrate esterases (CE), glycoside hydrolases (GH), glycosyltransferases (GT), and polysaccharide lyases (PL).

Selective pressure applied to the reactor led to a pH drop, and a significant variation over time in the abundance of specific microorganisms (bacterial MAGs) was observed (chi-square test P value, <0.001) (Figure 1A to C). At the same time, the community gene profiles were conserved (based on the profile of the clusters of orthologous groups, (COGs) (45)) (see Figure S1A in the supplementary material), as well as acetogenesis and methanogenesis pathways (based on the profile of the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology [KO]) (46) (Figures S1B and C in the supplementary material), including during the acidosis stage and showing the community functional redundancy. The hydrolytic potential of the whole community (separately investigated based on the GH family profile and the EC profile) was also maintained during the whole AD experiment, despite a slight decrease of the abundance of genes coding for hydrolytic enzymes during the acidosis stage (Figures 1E and F ; Figures S1D and E in the supplementary material). Three GH families were overrepresented (based on their metagenomic abundances and numbers of the different genes present) (Figures 1 and 2), including GH43 (average metagenomic abundance of 5.6% ± 1.1% of the total CAZyme gene content), GH2 (4.3% ± 0.3%), and GH109 (3.4% ± 0.5%). Other families such as GH28 (2.5% ± 0.3%), GH105 (2.1% ± 0.4%), GH3 (2.0% ± 2.0%), and GH13 (2.0% ± 0.2%) were also well represented. Despite their identical CAZy family annotations, different hydrolytic activities (EC category) were assigned to proteins representing a single

CAZy family (Figure 2C). Regarding their average metagenomic abundance, the hydrolytic portion of the metagenome was dominated by EC 3.2.1.8 (endo-1,4-β-xylanase; 8.2% ± 1.3% of all CAZyme-coding gene assigned to an EC category), EC 3.2.1.55 (α-L-arabinofuranosidase; 7.7% ± 4.4%), EC 3.2.1.37 (exo-1,4-β-xylosidase; 6.5% ± 0.8%), and EC 3.2.1.23 (β-galactosidase; 5.3% ± 0.5%) (Figure 1F). Similar to abundance of GHs, the average metagenomic abundance of CBM coding genes was highly correlated to the total gene copy numbers in the microbiome ($R2 = 0.91$, $P<0.001$) (Figure 2B). The set of putative CBMs fell into 36 families (Figure S1E in the supplementary material), with CBM50 being the most abundant. In total, 91 coding genes from CBM50 were identified, and 20.9% of them were associated with a GH domain in a single protein.

Out of the identified CAZyme genes, 65.3% were further assigned to the different reconstructed MAGs. BLAST analysis against the NCBI nonredundant protein database showed overall low identities of the identified CAZymes for most of the MAGs (Figure 2D). Roughly, 13.5% of predicted CAZymes showed high similarity (>95%) to the entries in the nonredundant protein database. This confirms the novelty of the identified enzymes. AA class seemed more conserved (higher sequence similarity), as previously demonstrated for the cow rumen metagenome (47). Interestingly, around 1.7% of all CAZyme coding genes were annotated as AAs. The AA genomic content of *Spirochaetes* was much higher than that for the other MAGs (5.4% ± 2.9% of the CAZyme-coding genes versus 2.5% ± 1.9% for *Firmicutes* and 0.6% ± 0.5% for *Bacteroidetes*).

3.2. *Bacteroidetes* may be favored in some anaerobic digesters owing to its polysaccharide hydrolytic potential as reflected by higher CAZy diversity and genomic content

We found that representatives of the phylum *Bacteroidetes* were among the most abundant MAGs across the digestion experiment (average metagenomic read abundance, 41.9% ± 8.4%), with MAG15 being dominant especially under optimal digestion conditions (Figure 1A). Interestingly, nearly half (49.2%) of the binned CAZyme-coding genes were assigned to *Bacteroidetes* MAGs. On average, they accounted for 8.8% ± 1.8% of their genome content, having higher frequency than the MAGs of *Firmicutes* (3.6% ± 1.1%) or *Spirochaetes* (2.9% ± 1.2%). Within

Bacteroidetes MAGs, 50.1% ± 5.6% of the CAZyme-coding genes were putative GHs, and 15.9% ± 4.0% were putative CEs, showing the high carbohydrate hydrolytic potential of this phylum (Figure 3A). The distribution into GHs and CEs was quite similar in *Firmicutes* (49.9% ± 12.5% and 12.6% ± 2% of all CAZyme-coding genes) and slightly higher in *Spirochaetes* (59.8% ± 4.7% and 13.9% ± 4.9%). In contrast, reconstructed "Candidatus *Cloacimonetes*" and *Synergistetes* MAGs had higher contents of GTs than GHs (39.7% ± 5.7% versus 21.1% ± 2.75% and 44.3% ± 19.0% versus 31.6% ± 14.7%, respectively). Interestingly, partially reconstructed MAG21, assigned to *Planctomycetes* (65.5% genome completeness) (Table S2 in the supplementary material), harbored multiple CAZyme-coding genes (representing 8.8% of its total genomic content), mainly assigned to GH127, GH2, GH43, GH5, and CBM35 families. The levels of diversity of assigned GH families and CBM families were the highest in the case of *Bacteroidetes* (Figures 3B and C). Around 23.2% of GH families and 24.1% of CBM families were specific to *Bacteroidetes* MAGs only, while roughly 8.7% of GH families and 17.1% of CBM families were absent from *Bacteroidetes* MAGs and present in the other MAGs. Even though the contents of GHs (as well as EC family assignments) differed between *Bacteroidetes* MAGs (Figure S2 in the supplementary material) at the phylum level, by adjusting the abundance of single populations, *Bacteroidetes* retained its hydrolytic potential (unchanged GH profile) during the whole experiment.

Bacteroidetes MAGs were highly enriched in sus-like genes, binning 94.3% of all sus-like conserved domains detected in the whole community metagenome (Figure 3A). On average, sus-like genes accounted for 2.4% ± 1.0% of the total gene content in *Bacteroidetes* MAGs, while they were almost absent from the other MAGs. In total, 49 putative PULs were predicted in *Bacteroidetes* MAGs. However, the final PUL number may be larger since its prediction is sensitive to the degree of the fragmentation of draft genomes (22).

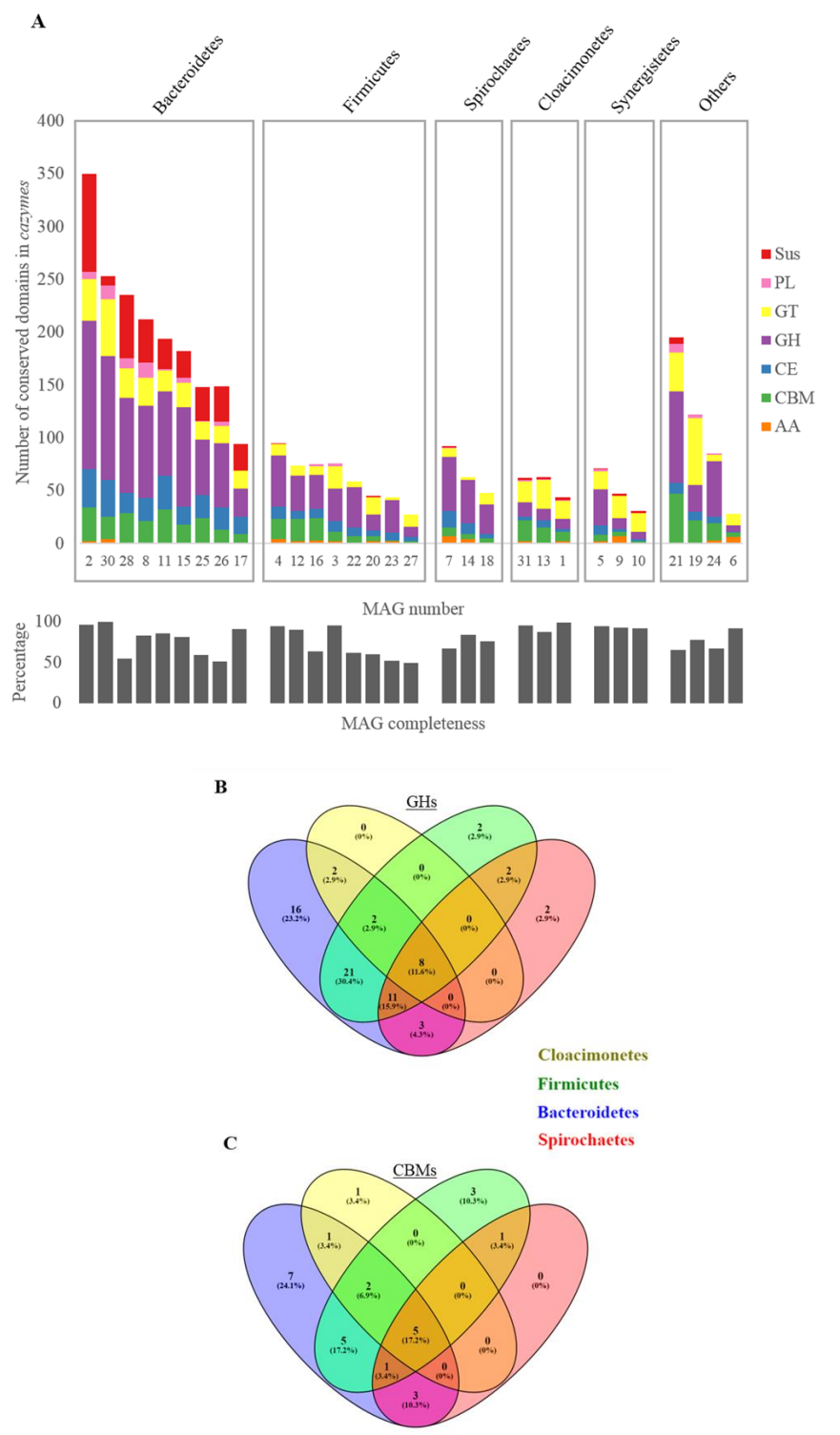

A
Bacteroidetes
Firmicutes
Spirochaetes
Cloacimonetes
Synergistetes
Others
Number of conserved domains in cazymes
400
350
300
250
200
150
100
50
0
Sus
PL
GT
GH
CE
CBM
AA
2 30 28 8 11 15 25 26 17
4 12 16 3 22 20 23 27
7 14 18
31 13 1
5 9 10
21 19 24 6
MAG number
Percentage
100
50
0
MAG completeness
B
GHs
16 (23.2%)
0 (0%)
2 (2.9%)
2 (2.9%)
0 (0%)
2 (2.9%)
2 (2.9%)
2 (2.9%)
0 (0%)
2 (2.9%)
21 (30.4%)
8 (11.6%)
0 (0%)
11 (15.9%)
0 (0%)
3 (4.3%)
Cloacimonetes
Firmicutes
Bacteroidetes
Spirochaetes
C
CBMs
7 (24.1%)
1 (3.4%)
3 (10.3%)
1 (3.4%)
0 (0%)
1 (3.4%)
2 (6.9%)
0 (0%)
0 (0%)
5 (17.2%)
5 (17.2%)
0 (0%)
1 (3.4%)
0 (0%)
3 (10.3%)

Figure 3: Characterization of the CAZy-coding genes, including sus-like genes, in the reconstructed metagenome assembled genomes (MAGs). **A** - Distribution of the different conserved domains assigned to CAZy-coding genes in the reconstructed MAGs. **B** and **C** - Venn diagram representation of the shared diversity of glycoside hydrolases (GHs) and carbohydrate binding modules (CBMs) in MAGs assigned to the four main bacterial phyla present in the reactor. For the description of MAGs, refer to Table S1 in the supplemental material. The different CAZy families represented include auxiliary enzymes (AA), carbohydrate-binding modules (CBM), carbohydrate esterases (CE), glycoside hydrolases (GH), glycosyltransferases (GT), and polysaccharide lyases (PL).

3.3. The diversity of CAZyme-coding genes in MAG15 might explain the metagenomic abundance of this species.

MAG15 was quite abundant during the whole anaerobic digestion process and especially under the optimal digestion conditions (Figure 1A, phase I). Using the database of 8,000 newly reconstructed MAGs from public metagenomic studies (48), we realized that MAG15 was absent from any other environment except our digesters. In addition, the novelty of putative CAZymes from this MAG was confirmed by their low similarities to other CAZyme proteins in the nonredundant protein database (Figure 4A). CAZyme-coding genes represented 9.1% of the genomic content of MAG15 (Table S2 in the supplementary material). Among the different CAZy domains identified, GH domains were the most dominant (representing 62.4% of the total CAZy domains in MAG15 genome). CBM and CE domains represented 11.4% and 6.7% of the CAZy domain content, respectively. The GH diversity of MAG15 was high and mainly represented by GH2, GH43, GH28, GH78, and GH3 families (Figure 4B), possibly allowing the species to cope with the different components of the sugar beet pulp. CBM67, CBM32, and CE6 domain-coding genes were also abundant in the MAG15 genome. Functionally assigned CAZyme-coding genes were represented by 31 EC classes with carbohydrate hydrolytic activity (Figure S2D in the supplementary material). Around 25 sus-like coding genes were identified in the genome of this new *Bacteroidetes* sp. and 76% of these genes were localized within PULs (Figure 4C). In total, seven PULs were identified in MAG15. However, due to the novelty of CAZyme-coding genes, only a few could have been assigned to an EC class, therefore preventing the in silico substrate specificity prediction of the identified PULs. PUL219 was the most complete and contained diverse CAZyme-coding genes

with putative functions predicted in silico to target acetylated glucomannan; therefore, it was further biochemically characterized.

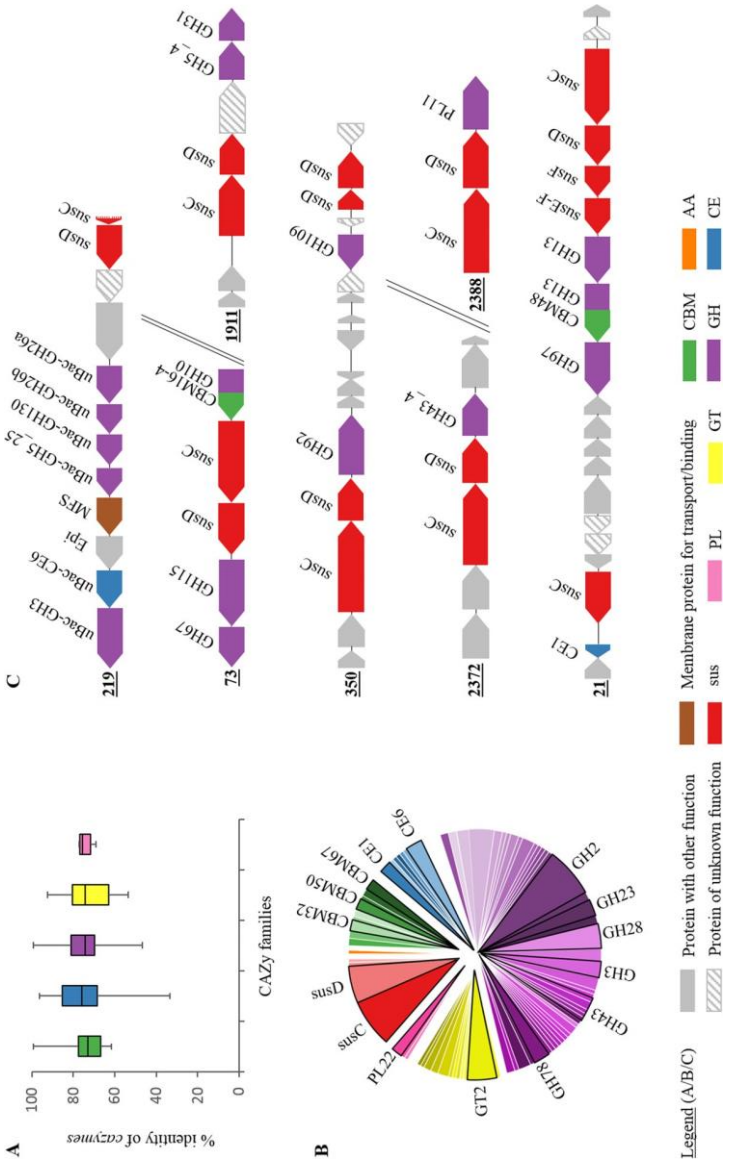


Figure 4: Characterization of the hydrolytic potential of a specific metagenome assembled genome (i.e. MAG15) assigned to an unknown *Bacteroidetes*. **A** - Box plot representations of the percentage identity of the CAZy-coding genes from MAG15 to proteins in the NCBI nonredundant protein database. **B** - Pie diagram showing the diversity of CAZy-coding genes in MAG15. **C** - Putative polysaccharide utilization loci (PULs) identified in the genome of MAG15. Numbers refer to the name of the identified PUL. Epi, N-acyl-D-glucosamine 2-epimerase; MFS, major facilitator superfamily. The legend refers to sus-like genes, auxiliary enzyme (AA), carbohydrate binding module (CBM), carbohydrate esterase (CE), glycoside hydrolase (GH), glycosyltransferase (GT), and polysaccharide lyase (PL).

3.4. Biochemical assays confirm predicted CAZy activities of six heterologously expressed proteins present in PUL219

PUL219 is composed of a putative pair of susC and susD homologs, one putative protein from the major facilitator superfamily (MFS), and six putative CAZymes belonging to five different families as well as a gene coding for a N-acyl-D-glucosamine 2-epimerase (Figure 4C). Following dbCAN annotation, uBac-GH26a and -b were assigned to the GH26 family, while uBac-GH130, uBac-GH5, uBac-CE6, and uBac-GH3 were assigned to the GH130, GH5, CE6, and GH3 families, respectively. Following the Hotpep analysis, uBac-GH26a and -b were functionally assigned to EC 3.2.1.78 (β-mannanase), uBac-GH130 was assigned to EC 2.4.1.281 (4-O-beta-D-mannosyl-D-glucose phosphorylase), uBac-GH5 was assigned to EC 3.2.1.4 (endoglucanase), and uBac-GH3 was assigned to EC 3.2.1.37 (exo-1,4-β-xylosidase). These genes were successfully cloned, and the corresponding proteins were expressed in *Escherichia coli* Rosetta(DE3) cells.

Results of the enzymatic activity assays and in silico signal peptide prediction are summarized in Table S3 in the supplementary material and Figure 5A. As correctly predicted, uBac-GH26a and uBac-GH26b showed capability of endo-action on galactomannan and glucomannan (Table S3 in the supplementary material; Figure 5A). They did not show any activity against the tested 4-nitrophenyl derivatives of monosaccharides, including 4-nitrophenyl α-D-mannopyranoside and 4-nitrophenyl β-D-mannopyranoside, showing no mannosidase activity with the tested substrates. Protein uBac-GH5 was active against 4-nitrophenyl β-D-cellobioside and was able to hydrolyze carboxymethyl cellulose (CMC). Moreover, it did not show any activity against 4-nitrophenyl β-D-glucopyranoside, confirming its predicted endoglucanase activity (EC 3.2.1.4). Finally, uBac-CE6 and uBac-GH3 were able to hydrolyze 4-nitrophenyl acetate and 4-nitrophenyl β-D-glucopyranoside, showing, respectively, esterase and β-glucosidase activities. The uBac-GH3 was shown to be a β-glucosidase (EC 3.2.1.21) rather than an exo-1,4-β-xylosidase (EC 3.2.1.37), as initially predicted by the Hotpep analysis. No activity was detected for uBAC-GH130 under the tested conditions and for the tested substrates. However, most of the characterized proteins assigned to this GH family exhibited phosphorylase activity (49). Additionally, the release of D-glucose, D-mannose, and acetic acid from acetylated konjac

glucomannan was further assessed by the use of the different enzymatic cocktails and/or cascade reactions (Figure 5A and Figure S3 in the supplementary material). As expected, acetic acid was released when the substrate was treated with esterase (uBac-CE6); however, initial deacetylation did not increase subsequent release of D-glucose and D-mannose compared to the level of non deacetylated konjac glucomannan. A larger amount of released D-glucose was measured when the enzymatic cocktail contained β-mannanase (cytoplasmic uBac-GH26a) and β-glucosidase (uBac-GH3). For a comparison, a smaller amount of released D-glucose residues was detected when the putatively periplasmic β-mannanase (uBac-GH26b) was combined in a cocktail with β-glucosidase. This finding suggests differential specificity of the two β-mannanases and indicates the formation of, respectively, longer and shorter mannooligosaccharides when the periplasmic and cytoplasmic isoforms are used, respectively. Similarly, a larger amount of released D-mannose was observed when the enzymatic cocktail contained the cytoplasmic form of the β-mannanase combined with the β-glucosidase (uBac-GH3) and independently of the prior substrate deacetylation (Figure S3 in the supplementary material). Additionally, using uBac-GH26a as a single enzyme resulted in mannose released from the treated konjac glucomannan. This result in combination with preliminary activity tests with synthetic substrates (e.g., no mannosidase activity detected using 4-nitrophenyl β-D-mannopyranoside) suggests that small-chain oligosaccharides from treated konjac glucomannan could be hydrolyzed by uBac-GH26a, releasing mannoses and shortened manno-oligosaccharides.

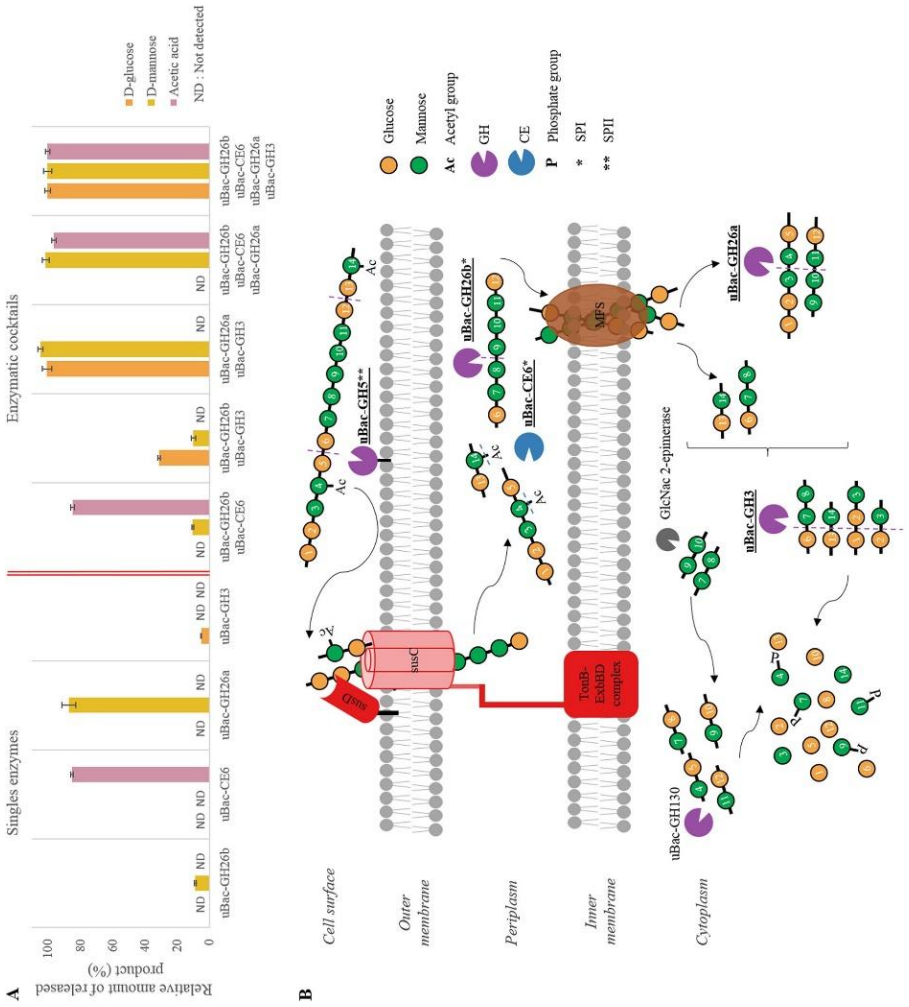


Figure 5: Characterization of a specific polysaccharide utilization locus (i.e. PUL219) toward acetylated glucomannan degradation. **A** - Release of D-glucose, D-mannose, and acetic acid by enzymatic hydrolysis of acetylated konjac glucomannan. Single enzymes as well as enzymatic cocktails were tested; for all the reactions, enzymatic hydrolysis of substrate lasted 1 h at 37°C. An enzymatic cocktail containing uBac-GH3, uBac-CE6, uBac-GH26a, and uBac-GH26b was taken as a reference to normalize results to 100%. **B** - Proposed mechanism of action of the putative acetylated glucomannan-targeting polysaccharide utilization locus PUL219, isolated from a new *Bacteroidetes* species. GlcNac 2-epimerase, N-acyl-D-glucosamine 2-epimerase; SPI, signal peptidase I predicted by LipoP, version 1.0; SPII, signal peptidase II predicted by LipoP, version 1.0. Enzymes biochemically characterized in this study are highlighted in bold and underlined.

4. Discussion

4.1. Functional redundancy of AD microbiome

Hydrolysis is recognized as a bottleneck of the AD process, especially in the case of highly lignified substrates (9). Additionally, an increased OLR of easily digestible substrates promotes fast hydrolysis, leading to VFA accumulation and the resulting process failure, known as acidosis. In a previous study, we showed that acidosis, promoted by increasing OLR of dry sugar beet pulp, affected the bacterial community structure, causing decreased methane production (Figures 1A to D) (4). During the acidosis phase, some CO_2 was still produced at a low rate compared to that under optimal conditions (linked to a lower total number of CAZyme-coding genes observed during acidosis in our study) (Figure 1), confirming the activity of hydrolytic bacteria at lower pH (50). Following the reestablishment of the optimal AD conditions in the reactor, increased hydrolysis was detected (measured by increased amount of the produced biogas) (Figure 1), suggesting that the hydrolytic readiness of the bacterial community was maintained during the process failure. In addition, we intended to investigate the microbiome functional redundancy with a special focus on its carbohydrate hydrolytic capacity. To that purpose, we applied metagenomics to seven samples taken at different stages of the experiment. Metagenomic results confirmed a significant variation in the abundances of individual bacteria over the experiment, based on the newly reconstructed MAGs (Figure 1). This is consistent with the previously performed 16S rRNA gene amplicon high-throughput characterization of the same microbial community (4). Functional redundancy, known to be widespread in different microbial environments (8), was previously shown at the metaproteome level for a steady-state anaerobic reactor (51). In our study, we showed that despite the differential genomic content of individual MAGs, a strong community functional redundancy was retained even during severe acidosis. This could explain why the reactor was able to quickly recover from acidosis and to promptly restart the production of biogas (functional profiles of genes involved in acetogenic and methanogenic pathways were largely retained as well) (see Figures S1B and C in the supplementary material). Taxonomic assignment of CAZyme-coding genes showed a shift in the bacterial population capable to hydrolyze the substrate along the

experiment (Figures 1B and C). Nevertheless, the genetic potential highlighted at the community-wide level by nearly identical CAZyme-coding gene profiles remained unchanged. This observation somehow fits the recently proposed premise in the context of the human gut of "function first, taxa second" (52). At the gene expression level, higher variability and sensitivity to perturbation than indicated by the content of the metagenome have previously been observed. Therefore, further metatranscriptomics study should complement the proposed functional redundancy of the AD microbiome by assessing its functional plasticity and, thus, its ability to adapt to perturbations by modulating the expression of the different genes.

4.2. Diversity of CAZymes

CAZymes are very important to the success of microbes in AD reactors fed with vegetal substrate as they are involved in the digestion of complex polysaccharides, which are particularly abundant in this environment (18). Here, we showed the usefulness of the complementary Hotpep analysis to the dbCAN-mediated CAZyme-coding gene discovery and classification. Indeed, within the abundant GH families (e.g., GH2, GH43, and GH3) (Figure 1E) different hydrolytic activities were detected on the basis of the assigned EC classes, e.g., α-L-arabinofuranosidase (EC 3.2.1.55), exo-1,4-β-xylosidase (EC 3.2.1.37), endo-1,4-β-xylanase (EC 3.2.1.8), β-glucosidase (EC 3.2.1.21), and β-galactosidase (EC 3.2.1.23) (Figure 2C). Still, the biochemical characterization of each new enzyme was necessary to provide correct insights into its saccharolytic capacity and substrate specificity. Sugar beet pulp is composed of diverse carbohydrates, including pectins enriched in arabinans and galactans, cellulose, and hemicelluloses, which comprise small amounts of glucomannans and xyloglucans (53). Accordingly, CAZymes assigned to the most abundant EC categories were predicted to be mainly involved in the hydrolysis of hemicellulosic material and especially (arabino)xylan, (arabino)galactan, and xyloglucan-enriched hemicelluloses. Thus, the ability of the microbiome to hydrolyze sugar beet pulp could be further linked to the abundance of putative β-galactosidases (EC 3.2.1.23) as well as α-L-arabinofuranosidases (EC 3.2.1.55), which were shown to be largely involved in galactan and arabinan degradation (54, 55). Sugar beet pulp constituent-targeting CBMs were also abundant, including pectin-targeting CBM32 or CBM67 that were shown to bind to polygalacturonic acid or α-L-rhamnoside, respectively (56, 57).

The capacity of the microbiome to use different carbon sources was also confirmed by the presence of other CBMs, with CBM50 and CBM56 targeting chitin and peptidoglycan (e.g., fungal and bacterial cell wall components, respectively) being the most abundant. Moreover, family GH109 has often been recorded in metagenomic studies of anaerobic digesters (58), including in our reactor. This family is known to express the α-N-acetylgalactosaminidase activity (EC 3.2.1.49), possibly targeting acetylgalactosamine, which is a component of bacterial cell walls (48). On one hand, the abundance of genes assigned to this family, together with the presence of CBM50 and CBM56 domains, might be linked to the expected high turnover rates of bacterial biomass resulting from high competition in this environment (59). On the other hand, only a few GH109-classified CAZyme-coding genes were assigned to the respective EC 3.2.1.49 category (0.4% ± 0.3% of metagenomic abundance) (Figure 2C). Therefore, further investigation is required to confirm the specific (hydrolytic) activity of this family of proteins. The presence of AA coding genes, including AA2 containing class II lignin-modifying peroxidases, indicates the potential to deconstruct lignin by some anaerobic digester microbes. Demethylation of lignin was suggested for some *Spirochaetes* isolated from the termite gut (60), which is in accordance to our *Spirochaetes* MAGs being enriched in AA coding genes.

4.3. *Bacteroidetes* and their PULs

Similarly to findings of previous studies of AD microbiomes (11, 61, 62), *Bacteroidetes* MAGs were shown to be among of the most dominant in the reconstructed metagenome representing the digester microbial community. Indeed, higher genomic content and functional diversity of hydrolytic genes (GHs and ECs) in *Bacteroidetes* MAGs (Figure 3A and Figure S4 in the supplementary material) could ensure digestibility of a wide range of substrates, thus favoring their abundance in AD reactors under certain environmental conditions (11). Individual *Bacteroidetes* MAGs were characterized with distinct CAZyme-coding gene functional profiles suggesting niche differentiation (different nutritional requirements) of the different species (Figure S2 in the supplementary material). However, at the phylum level the *Bacteroidetes* population, by regulating the abundance of single individuals, was capable of ensuring the stability of gene functional profiles, thus largely contributing to the community-wide functional redundancy. *Bacteroidetes* and some other bacteria

(e.g., *Firmicutes* and representatives of *Spirochaetes*, *Lentisphaerae*, *Planctomycetes*, and *Thermotogae*) that harbor multiple CAZyme-coding genes in their genomes (Table S2 in the supplementary material) seem responsible for complex carbohydrate degradation (15). Interestingly, *Planctomycetes* MAG21 harbors a CAZy profile similar to that of *Bacteroidetes* MAGs with a relatively high CAZyme-coding gene content (Figure 3A). Even though the role of this bacterium in anaerobic digestion has not been widely discussed, previous studies suggest the involvement of *Planctomycetes* in degradation of chitin and cellulose in agricultural soil (63) and humus deconstruction in the termite gut (64). These observations might indicate a similar hydrolytic potential to *Bacteroidetes* and the importance of *Planctomycetes* in the anaerobic digestion process (due to its ability to deconstruct recalcitrant organic matter). In contrast, "Ca. *Cloacimonetes*" and *Synergistetes* have different CAZy profiles, mainly enriched in GTs, and seem of less importance for carbohydrate hydrolysis. Nevertheless, *in situ* hybridization using an iodine-labeled oligonucleotide probe combined with high-resolution nanometer scale secondary ion mass spectrometry (SIMS) supported the role of "Ca. *Cloacimonetes*" in the fermentative digestion of cellulose (65). Owing to its status of a candidate phylum and to the recently proposed diversity of "Ca. *Cloacimonetes*" in the different full-scale AD reactors (11), a broad range of functional diversity may be expected within this phylum.

MAG15 assigned to an unknown *Bacteroidales* was characterized with relatively high community abundance in the reactor during the whole AD experiment (Figure 1A). According to its CAZy profile (Figures 4B and C), it showed potential to digest a broad range of polysaccharides present in sugar beet pulp. CAZyme-coding genes assigned to the family GH78 known to have α-L rhamnosidase activity (EC 3.2.1.40) acting on rhamnogalacturonan, a component of pectins (55), were present in multiple copies in its genome. In addition, the presence of numerous CAZyme-coding gene modules from the CBM67 family having α-L-rhamnose binding activity (57) confirms the potential of this bacterium to target pectins (Figure 4B). The presence of (hetero)xylan targeting CAZyme genes, including the putative enzymes attacking the backbone (endoxylanases and xylosidases assigned to e.g. families GH2, -3, -10, and -43), as well as arabinases (e.g. GH2 and 43) and galactosidases (e.g. GH2), confirms the ability of MAG15 to deconstruct different forms of xylans (e.g. arabinoxylan and

glucuronoarabinoxylan). Many novel *Bacteroidetes* were shown to contain various sets of PULs predicted to be involved in complex hydrolysis of diverse carbohydrates (20), changing our understanding of polysaccharide metabolism inside this phylum. The partially reconstructed MAG15 contains multiple PULs, including PUL73 predicted to target xylan, and the starch-targeting PUL21. PUL219, containing the highest diversity of putative CAZyme-coding genes, was additionally experimentally validated to target (acetylated) glucomannan (see section results). Glucomannan is a polysaccharide composed of β-D-1,4-linked mannose and glucose monomers (possible combinations include glucose-mannose, glucose-glucose, mannose-glucose, and mannose-mannose linkages), with side chain acetyl groups (Figure 5B) (66, 67). Previously biochemically validated PULs included a galactomannan-targeting PUL from a marine *Bacteroidetes* (68) and a cellulose-degrading one from a rumen-isolated bacterium (69). Here, we propose a model of PUL-assisted acetylated glucomannan degradation for a *Bacteroidetes* species (Figure 5B). This model was constructed based on the activity tests performed in this study using purified proteins and complementary in silico predictions, as described below. According to the predicted presence of a lipoprotein signal peptidase II and a cleavage site within uBac-GH5 sequence, this endoglucanase is most probably attached to the outer membrane, extracellularly digesting acetylated glucomannans and releasing acetylated glucomannan oligosaccharides (70, 71). As susD is attached to the outer membrane, it binds the resulting substrate and directs it to the susC transporter, which passes it to the periplasm (72). A signal peptidase I and a cleavage site were in silico predicted in both uBac-GH26b and uBac-CE6, suggesting their migration from the cytoplasm to the periplasm (73). In the periplasm, β-mannanase (uBac-GH26b) further hydrolyzes the acetylated glucomannan oligomers while acetyl groups are removed by the action of an esterase (uBac-CE6). Even shorter oligosaccharides may now be transported through the inner membrane with the help of the MFS (74). In the cytoplasm, short chains of glucomannan are further digested by uBac-GH26a (second β-mannanase), possibly releasing mono-, di-, and trisaccharides (75). These oligosaccharides are further hydrolyzed by uBac-GH3 (β-glucosidase), releasing glucose, mannose, and mannobiose. N-Acyl-D-glucosamine 2-epimerase transforms mannobiose into β-D-mannosyl-(1¡4)-D-glucose (MannosylGlu). Hydrolysis of MannosylGlu to glucose and mannose-1-phosphate is mediated by uBac-GH130 (presumed to be the 4-O-beta-

D-mannosyl-D-glucose phosphorylase) (76, 77). A similar galactomannan-targeting PUL from *Bacteroides ovatus* has been previously characterized (78), including two β-mannanases from the GH26 family (BoMan26A and BoMan26B). In that study, BoMan26B was shown to release longer chain oligosaccharides than BoMan26A, similarly to uBac-GH26b and uBac-GH26a as proposed in our model (Figure 5). Phylogenetic analysis of several mannanases (including previously characterized BoMan26A and BoMan26B) confirmed the presence of two distinct mannanase clusters inside the GH26 family, showing the differential clustering of uBac-GH26a and uBac-GH26b, which are closely related to BoMan26A and BoMan26B, respectively (Figure S5 in the supplementary material).

Further characterization of novel PULs should bring even more interesting insights into the complex carbohydrate metabolism in *Bacteroidetes*. Specific substrate targeting ready-to-use enzymatic cocktails naturally selected in PULs might be of high interest to the biotechnology sector as well, especially for developing biomass-based green chemistry. For example, acetylated glucomannan is the main hemicellulosic component of the secondary cell walls in soft woods (79), which are common substrates for pulp and paper production. A nature-evolved cocktail of acetylated glucomannan-targeting enzymes, as characterized in e.g. PUL219, might be an interesting alternative to the currently used chemical pretreatments applied in these industries.

5. Conclusion

Our MG analysis applied to an anaerobic digestion microbiome fed with sugar beet pulp, revealed a large carbohydrate hydrolytic potential (more than 4000 CAZymes-coding genes detected) and redundancy (CAZyme profile unchanged), despite the evolution of the bacterial community composition related to a change in the AD process towards acidosis. The presence of PULs in one of the most abundant *Bacteroidetes* MAG was demonsrated. A set of CAZymes identified in a PUL, was heterologously produced, and tested against a diversity of substrates. The *in vitro* activities confirmed most activities predicted *in silico*. The set of results allowed to propose a mode of action of this PUL, identified as targeting acetylated glucomannans.

Acknowledgements

We thank Dominika Klimek, Sebastien Lemaigre, and Stephanie Giusti-Miller for their technical support and Kjell Sergent for his expertise in protein identification analysis. This work was supported by the Luxembourg National Research Fund, as follows: by FNR CORE 2011 project GASPOP (C11/SR/1280949: Influence of the Reactor Design and the Operational Parameters on the Dynamics of the Microbial Consortia Involved in the Biomethanation Process) and by FNR CORE 2014 project OPTILYS (C14/SR/8286517: Exploring the higher termite lignocellulolytic system to optimize the conversion of biomass into energy and useful platform molecules).

References

1. B. F. Pain, R. Q. Hepherd, Anaerobic digestion of livestock wastes in *Anaerobic Digestion of Farm Waste. Proceedings of Meeting, NIRD, Reading, UK,* (1985), p. pp.9-14.
2. B. K. Ahring, "Perspectives for anaerobic digestion" in *Biomethanation I*, 81st Ed., B. Ahring, *et al.*, Eds. (Springer, Berlin, Heidelberg, 2003), pp. 1–30.
3. S. Lemaigre, *et al.*, Transfer of a static PCA-MSPC model from a steady-state anaerobic reactor to an independent anaerobic reactor exposed to organic overload. *Chemom. Intell. Lab. Syst.* **159**, 20–30 (2016).
4. X. Goux, *et al.*, Microbial community dynamics in replicate anaerobic digesters exposed sequentially to increasing organic loading rate, acidosis, and process recovery. *Biotechnol. Biofuels* **8** (2015).
5. M. H. Gerardi, *The microbiology of anaerobic digesters*, M. H. Gerardi, Ed., 1st Ed. (John Wiley & Sons, Inc., 2003).
6. S. Lerm, *et al.*, Archaeal community composition affects the function of anaerobic co-digesters in response to organic overload. *Waste Manag.* **32**, 389–399 (2012).
7. X. Goux, *et al.*, Microbial community dynamics in replicate anaerobic digesters exposed sequentially to increasing organic loading rate, acidosis, and process recovery. *Biotechnol. Biofuels* **8**, 1–18 (2015).
8. S. Louca, *et al.*, Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* **2**, 936–943 (2018).
9. C. Álvarez, F. M. Reyes-Sosa, B. Díez, Enzymatic hydrolysis of biomass from wood. *Microb. Biotechnol.* **9**, 149–156 (2016).
10. P. Coutinho, B. Henrissat, "Carbohydrate-active enzymes: an integrated database approach" in *Recent Advances in Carbohydrate Bioengineering*, H. Gilbert, G. Davies, H. Henrissat, B. Svensson, Eds. (Cambridge: The Royal Society of Chemistry, 1999), pp. 3–12.
11. M. Calusinska, *et al.*, A year of monitoring 20 mesophilic full - scale bioreactors

reveals the existence of stable but different core microbiomes in bio - waste and wastewater anaerobic digestion systems. *Biotechnol. Biofuels* **11** (2018).
12. L. Sun, P. B. Pope, V. G. H. Eijsink, A. Schnürer, Characterization of microbial community structure during continuous anaerobic digestion of straw and cow manure. *Microb. Biotechnol.* **8**, 815–827 (2015).
13. X. Goux, M. Calusinska, M. Fossépré, E. Benizri, P. Delfosse, Start-up phase of an anaerobic full-scale farm reactor - Appearance of mesophilic anaerobic conditions and establishment of the methanogenic microbial community. *Bioresour. Technol.* **212**, 217–226 (2016).
14. S. Lemaigre, *et al.*, Potential of multivariate statistical process monitoring based on the biogas composition to detect free ammonia intoxication in anaerobic reactors. *Biochem. Eng. J.* **140**, 17–28 (2018).
15. S. Campanaro, L. Treu, P. G. Kougias, G. Luo, I. Angelidaki, Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water Res.* **140**, 123–134 (2018).
16. O. Svartström, *et al.*, Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation. *ISME J.* **11**, 2538–2551 (2017).
17. I. Vanwonterghem, P. D. Jensen, K. Rabaey, G. W. Tyson, Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion. *Environ. Microbiol.* **18**, 3144–3158 (2016).
18. P. G. Kougias, *et al.*, Spatial distribution and diverse metabolic functions of lignocellulose-degrading uncultured bacteria as revealed by genomecentric metagenomics. *Appl. Environ. Microbiol.* **84**, 1–14 (2018).
19. G. Gefen, M. Anbar, E. Morag, R. Lamed, E. a. Bayer, Enhanced cellulose degradation by targeted integration of a cohesin-fused -glucosidase into the Clostridium thermocellum cellulosome. *Proc. Natl. Acad. Sci.* **109**, 10298–10303 (2012).
20. J. M. Grondin, K. Tamura, G. Déjean, D. W. Abbott, H. Brumer, Polysaccharide Utilization Loci: Fuelling microbial communities. *J. Bacteriol.* **199**, JB.00860-16 (2017).
21. M. K. Bjursell, E. C. Martens, J. I. Gordon, Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. *J. Biol. Chem.* **281**, 36269–36279 (2006).
22. N. Terrapon, V. Lombard, H. J. Gilbert, B. Henrissat, Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **31**, 647–655 (2015).
23. E. C. Martens, N. M. Koropatkin, T. J. Smith, J. I. Gordon, Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. *J. Biol. Chem.* **284**, 24673–24677 (2009).
24. K. L. Anderson, A. A. Salyers, Biochemical Evidence that Starch Breakdown by Bacteroides thetaiotaomicron Involves Outer Membrane Starch-Binding Sites and Periplasmic Starch-Degrading Enzymes. **2**, 3192–3198 (1989).

25. A. Rogowski, *et al.*, Glycan complexity dictates microbial resource allocation in the large intestine. *Nat. Commun.* **6**, 7481 (2015).
26. M. Zhang, *et al.*, Xylan utilization in human gut commensal bacteria is orchestrated by unique modular organization of polysaccharide-degrading enzymes. *Proc. Natl. Acad. Sci.* **111**, E3708–E3717 (2014).
27. K. Kawaguchi, T. Senoura, S. Ito, The mannobiose - forming exo - mannanase involved in a new mannan catabolic pathway in Bacteroides fragilis. 17–23 (2014).
28. S. K. Reddy, *et al.*, A β-mannan utilization locus in Bacteroides ovatus involves a GH36 α-galactosidase active on galactomannans. *FEBS Lett.* **590**, 2106–2118 (2016).
29. S. L. La Rosa, *et al.*, Wood-Derived Dietary Fibers Promote Beneficial Human Gut Microbiota. *mSphere* **4**, 1–16 (2019).
30. K. Tang, Y. Lin, Y. Han, N. Jiao, Characterization of potential polysaccharide utilization systems in the marine Bacteroidetes Gramella flava JLT2011 using a multi-omics approach. *Front. Microbiol.* **8**, 1–13 (2017).
31. J. Larsbrink, *et al.*, A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
32. B. Broeksema, *et al.*, ICoVeR – an interactive visualization tool for verification and refinement of metagenomic bins. *BMC Bioinformatics* **18**, 233 (2017).
33. N. Segata, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes (2013) https:/doi.org/10.1038/ncomms3304.PhyloPhlAn.
34. D. Hyatt, *et al.*, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11** (2010).
35. Y. Yin, *et al.*, DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, 445–451 (2012).
36. N. Terrapon, *et al.*, PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res.* **46**, 677–683 (2017).
37. P. K. Busk, B. Pilgaard, M. J. Lezyk, A. S. Meyer, L. Lange, Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics* **18**, 1–9 (2017).
38. V. M. Markowitz, *et al.*, IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, 115–122 (2012).
39. R. K. Aziz, *et al.*, The RAST Server : Rapid Annotations using Subsystems Technology. **15**, 1–15 (2008).
40. M. B. Jaffee, B. Imperiali, Optimized protocol for expression and purification of membrane-bound PglB, a bacterial oligosaccharyl transferase. *Protein Expr. Purif.* **89**, 241–250 (2013).
41. J. Tiralongo, A. Maggioni, "The Targeted Expression of Nucelotide Sugar Transporters to the E.coli Inner Membrane" in *Heterologous Gene Expression in E.Coli*, (2010), pp. 237–249.
42. A. S. Juncker, *et al.*, Prediction of lipoprotein signal peptides in Gram-negative

bacteria. *Protein Sci.* **12**, 1652–1662 (2003).
43. C. Rouland, A. Civas, J. Renoux, F. Petek, Synergistic activities of the enzymes involved in cellulose degradation purified from Macrotermes mülleri and from its symbiotic fungus Termitomyces sp. *Comp. Biochem. Physiol.* **91** (1988).
44. V. Flari, M. Matoub, C. Rouland, Purification and characterization of a ß-mannanase from the digestive tract of the edible snail Helix lucorum L. *Carbohydr. Res.* **275**, 207–213 (1995).
45. M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Expanded Microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2015).
46. M. Kanehisa, Y. Sato, K. Morishima, BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
47. R. D. Stewart, *et al.*, Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 1–11 (2018).
48. D. H. Parks, *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2** (2017).
49. W. Saburi, Functions , structures , and applications of cellobiose 2-epimerase and glycoside hydrolase family 130 mannoside phosphorylases. *Biosci. Biotechnol. Biochem.* **80**, 1294–1305 (2016).
50. K. Wang, J. Yin, D. Shen, N. Li, Anaerobic digestion of food waste for volatile fatty acids (VFAs) production with different types of inoculum: Effect of pH. *Bioresour. Technol.* **161**, 395–401 (2014).
51. A. Joyce, *et al.*, Linking microbial community structure and function during the acidified anaerobic digestion of grass. *Front. Microbiol.* **9**, 1–13 (2018).
52. A. Heintz-Buschart, P. Wilmes, Human Gut Microbiome: Function Matters. *Trends Microbiol.* **26**, 563–574 (2018).
53. A. Zykwinska, C. Rondeau-mouro, C. Garnier, Alkaline extractability of pectic arabinan and galactan and their mobility in sugar beet and potato cell walls. **65**, 510–520 (2006).
54. H. Watzlawick, K. Morabbi Heravi, J. Altenbuchner, Role of the ganSPQAB Operon in Degradation of Galactan by Bacillus subtilis. **198**, 2887–2896 (2016).
55. D. Wefers, *et al.*, Enzymatic Mechanism for Arabinan Degradation and Transport in the Thermophilic Bacterium Caldanaerobius polysaccharolyticus. **83**, 1–21 (2017).
56. D. W. Abbott, S. Hrynuik, A. B. Boraston, Identification and Characterization of a Novel Periplasmic Polygalacturonic Acid Binding Protein from Yersinia enterolitica. *J. Mol. Biol.* **367**, 1023–1033 (2007).
57. Z. Fujimoto, *et al.*, The structure of a Streptomyces avermitilis α-L-Rhamnosidase reveals a novel carbohydrate-binding module CBM67 within the six-domain arrangement. *J. Biol. Chem.* **288**, 12376–12385 (2013).
58. I. Maus, *et al.*, Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization

of bacterial and archaeal isolates. *Biotechnol. Biofuels*, 1–28 (2016).
59. A. M. Alessi, *et al.*, Defining functional diversity for lignocellulose degradation in a microbial community using multi-omics studies. *Biotechnol. Biofuels* **11**, 1–16 (2018).
60. G. R. Esenther, T. K. Kirk, Catabolism of aspen sapwood in *Reticulitermes flavipes*. *Ann. Entomol. Soc. Am.* **67**, 989–991 (1974).
61. Y. F. Li, *et al.*, Comparison of the microbial communities in solid-state anaerobic digestion (SS-AD) reactors operated at mesophilic and thermophilic temperatures. *Appl. Microbiol. Biotechnol.* **99**, 969–980 (2014).
62. J. Guo, *et al.*, Dissecting microbial community structure and methane-producing pathways of a full-scale anaerobic reactor digesting activated sludge from wastewater treatment by metagenomic sequencing. *Microb. Cell Fact.* **14**, 1–11 (2015).
63. A. S. Wieczorek, S. A. Hetz, S. Kolb, Microbial responses to chitin and chitosan in oxic and anoxic agricultural soil slurries. *Biogeosciences* **11**, 3339–3352 (2014).
64. T. Köhler, U. Stingl, K. Meuser, A. Brune, Novel lineages of Planctomycetes densely colonize the alkaline gut of soil-feeding termites (Cubitermes spp.). *Environ. Microbiol.* **10**, 1260–1270 (2008).
65. R. D. Limam, *et al.*, Members of the uncultured bacterial candidate division WWE1 are implicated in anaerobic digestion of cellulose. *Microbiol. Open* **3**, 157–167 (2014).
66. C. Zhang, J. Chen, F. Yang, Konjac glucomannan , a promising polysaccharide for OCDDS. *Carbohydr. Polym.* **104**, 175–181 (2014).
67. A. Mikkelson, H. Maaheimo, T. K. Hakala, Hydrolysis of konjac glucomannan by Trichoderma reesei mannanase and endoglucanases Cel7B and Cel5A for the production of glucomannooligosaccharides. *Carbohydr. Res.* **372**, 60–68 (2013).
68. J. Chen, *et al.*, Alpha- and beta-mannan utilization by marine *Bacteroidetes*. *Environ. Microbiol.* **20**, 4127–4140 (2018).
69. A. E. Naas, *et al.*, Do rumen *Bacteroidetes* utilize an alternative mechanism for cellulose degradation? *MBio* **5**, 4–9 (2014).
70. M. J. Temple, *et al.*, A Bacteroidetes locus dedicated to fungal 1,6--glucan degradation: Unique substrate conformation drives specificity of the key endo-1,6--glucanase. *J. Biol. Chem.* **292**, 10639–10650 (2017).
71. W. R. Zückert, Secretion of Bacterial Lipoproteins: Through the Cytoplasmic Membrane, the Periplasm and Beyond. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1509–1516 (2014).
72. J. A. Shipman, J. E. Berleman, A. A. Salyers, Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. *J. Bacteriol.* **182**, 5365–5372 (2000).
73. A. P. Pugsley, The complete general secretory pathway in gram-negative bacteria. *Microbiol. Rev.* **57**, 50–108 (1993).
74. A. S. Tauzin, *et al.*, Functional characterization of a gene locus from an uncultured gut Bacteroides conferring xylo-oligosaccharides utilization to Escherichia coli.

Mol. Microbiol. **102**, 579–592 (2016).
75. G. C. Pradeep, *et al.*, An extremely alkaline mannanase from Streptomyces sp. CS428 hydrolyzes galactomannan producing series of mannooligosaccharides. *World J. Microbiol. Biotechnol.* **32**, 1–9 (2016).
76. T. Senoura, *et al.*, New microbial mannan catabolic pathway that involves a novel mannosylglucose phosphorylase. *Biochem. Biophys. Res. Commun.* **408**, 701–706 (2011).
77. R. Kawahara, *et al.*, Metabolic mechanism of mannan in a ruminal bacterium, Ruminococcus albus, involving two mannoside phosphorylases and cellobiose 2-epimerase: Discovery of a new carbohydrate phosphorylase, β-1,4-mannooligosaccharide phosphorylase. *J. Biol. Chem.* **287**, 42389–42399 (2012).
78. V. Bagenholm, *et al.*, Galactomannan catabolism conferred by a polysaccharide utilization locus of Bacteroides ovatus: Enzyme synergy and crystal structure of a β-mannanase. *J. Biol. Chem.* **292**, 229–243 (2017).
79. P. M.-A. Pawar, S. Koutaniemi, M. Tenkanen, E. J. Mellerowicz, Acetylation of woody lignocellulose: significance and regulation. *Front. Plant Sci.* **4**, 118 (2013).

Chapter 3

Multi-omics applied to bioprospecting carbohydrate active enzymes in the anaerobic digestion process: focus on α-L-arabinofuranosidases and feruloyl esterases

In preparation for publication

Authors: Bertucci M., Calusinska M., Goux X., Lemaigre S., Wilmes, P., Gerin P. A., Delfosse P.

Personal contribution:

- Full analysis of metagenomics and metatranscriptomics datasets
- Redaction of the full chapter including the comments and corrections of the co-authors

In the previous chapter (Chapter 2), metagenomics investigation was applied to an anaerobic digestion microbiome exposed to acidosis. The main outputs were: (1) the assessment of the carbohydrate hydrolytic potential and functional redundancy of the microbiome (2) an insight in the bacterial strategies to deconstruct lignocellulosic biomass (3) a model for the mode of action of an acetylated glucomannan-targeting PUL and (4) the suggestion of including accessory enzymes in cocktails to be used in the biorefinery sector.

In the present chapter (Chapter 3), metagenomics was complemented with metatranscriptomics to investigate an anerobic digestion microbiome fed with sugar beet pulp. The main goal of this study was to (1) define the carbohydrate hydrolytic potential of the bacterial community (metagenomics) (2) identify key bacterial players (evidenced by a combined metagenomic and metatranscriptomic approach) involved in LCB deconstruction and (3) characterize gene clusters containing carbohydrate active enzymes (especially α-L-arabinofuranosidases and feruloyl esterases, identified as key enzymes involved in lignin detachment). This is a contribution to the bioprospection for enzymatic and bacterial resources available in such an environment.

Abstract

Being the most abundant and renewable natural resource on Earth, plant biomass is considered as a source of fermentable sugars and chemicals for many industrial sectors, including biofuel, pulp and paper as well as food and feed. However, its composition (sometimes highly lignified) confers its recalcitrant property. On the one hand, man-made processes for plant biomass deconstruction include environmentally non-friendly thermal and chemical pre-treatments. On the other hand, biomass natural degradation by living organisms requires an arsenal of enzymes, comprising carbohydrate active enzymes (CAZymes). Recently, researchers have discovered that microbes, more specifically *Bacteroidetes*, harbor in their genomes polysaccharide utilization loci (PULs), i.e. clusters of co-expressed genes targeting a specific component of the plant lignocellulose. These nature-made complexes could be an alternative to the commercially already existing enzymatic cocktails. In this discovery-driven study, we apply high-throughput sequencing technologies (i.e. metagenomics and metatrancriptomics) to the microbiome of an anerobic digester. We identified the main actors of the biomass hydrolysis, i.e. *Bacteroidetes* and *Chloroflexi*. They were identified as the two main phyla contributing to the abundance of CAZymes, and more specifically glycoside hydrolases (GHs). In total, 2745 GHs were assigned to *Bacteroidetes* (87,2% were expressed in the investigated anaerobic digeston process) and 2436 GHs were assigned to *Chloroflexi* (56.4% expressed). α-L-arabinofuranosidases and feruloyl esterases have been identified. They are accessory enzymes disrupting the links between arabinoxylan and lignin. They were shown to be involved in putative PULs targeting (glucurono)arabinoxylan. As lignin is the most recalcitrant component of lignocellulosic biomass, this discovery could be the starting point for the selection of accessory enzymes from PULs to valorize their potential enhancing biomass hydrolysis in the biorefinery sector.

Keywords: biorefineries, carbohydrate hydrolytic potential, α-L-arabinofuranosidases, feruloyl esterases, biomass recalcitrance, multi-omics analysis

1. Introduction

There are growing appeals for the green industry sector, involving the use of the most abundant and renewable natural resource, i.e. plant biomass. Plant cell wall is mainly composed of cellulose, hemicellulose, pectin and lignin (1, 2). Enzymatic hydrolysis of biomass could offer several benefits over chemical and thermal treatments in terms of sustainability concerns and production cost, and especially substrate selectivity. Actually, the production of enzymes is a central ques to the growing biotechnology industry (3). Regarding the complex nature of lignocellulose polysaccharides, their complete degradation requires a large enzymatic arsenal. Most of the commercially available enzymatic cocktails contain a broad mix of so-called "core enzymes" (targeting the backbone of a specific plant polymer, e.g. cellulose, mannan, xylan) of diverse origin (mainly fungal), and accessory debranching (removing side chains) and/or mobilising (breaking bonds between diverse plant polymers) enzymes, together with other non-enzymatic components (4, 5). In this sense, xylanases and cellulases are the main lignocellulose-degrading enzymes used at industrial scale (6). Natural complexes (clusters of co-expressed genes) of carbohydrate active enzymes (CAZymes), encoded in microbial genomes and specifically targeting different biomass components, might provide a future state-of-the-art solution for the emerging bio-refineries sector. With the advancement of the high-throughput sequencing and heterologous enzyme production technologies, their discovery and characterisation from different microbial environments are now easier than ever before. Indeed, bioprospecting (discovery and characterisation with the view of potential commercialisation of new products based on biological resources) for novel enzymes from e.g. herbivores rumen, termite gut, anaerobic reactors etc. offers an alternative to commercial enzyme preparations (4). In *Bacteroidetes*, CAZymes form clusters known as polysaccharide utilization loci (PULs)(7). Their particularity lies in the presence of a pair of specific carbohydrate transporters, i.e. *Sus*-like genes, encoding for a TonB-dependent transporter (SusC) and a cell-surface glycan-binding protein (SusD)(8).

Anaerobic digestion (AD) is a nature-inspired process optimized by humans and leading to biomethame production through biomass degradation. Previously, the metagenomic investigation of a sugar beet pulp acclimated microbiome has shown

that AD can be considered as a reservoir of CAZymes coding genes and provided us great insight into substrate hydrolytic deconstruction by microbes. Complete PULs were revealed, specifically targeting different fractions of SBP (9) and chapter 3. Coupling metatranscriptomics to metagenomics, could give additional informations about the genes expressed, and therefore the PULs that might be of interest regarding their substrate specificity.

Particularly, accessory enzymes are poorly investigated, in comparision to hemicellulases and cellulases. However, they might be of high interest for the green biotechnology sector. Indeed, branched component build the link between the different polysaccharides, and targeting these compound might lead to an increased accessibility of the "core-enzymes", and therefore, increase the efficiency of biomass degradation (10). In particular, in grasses, ferulate is linked to arabinose and forms covalent linkages between arabinoxylan chains as well as between arabinoxylan and lignin components, therefore limiting the accessibility of enzymes to their substrate (11). Therefore, the application of novel α-L-arabinofuranosidases (EC 3.2.1.55) and feruloyl esterases (EC 3.1.1.73) might be of interest to enhance biomass hydrolysis, and these enzymes are often components of hydrolytic enzyme mixes. Specifically, α-L-arabinofuranosidases (ARAFs) hydrolyse arabinose side-chain compounds in arabinoxylan (12). Next to ARAF, feruloyl esterase (FAE) is catalyzing the hydrolysis of ester linkages between ferulic acid (which can be linked to lignin) and arabinoxylan, therefore disrupting the bonds between lignin and the polysaccharide (13). Therefore, FAEs play an important role in promoting the bioaccessibility of hemicellulases to their substrate (14). Additionally, several studies have also shown that the additional use of FAEs during enzymatic degradation of lignocellulosic biomass enhanced the overall saccharification process (15, 16). For all these concerns, in this discovery-driven chapter, we harvest the carbohydrate hydrolytic potential of an anaerobic digester by exploring with metagenomics and metatranscriptomics approaches focused on PULs. ARAFs and FAEs were more specifically investigated.

2. Material and methods

2.1. Anaerobic digestion experiment, samples and sequencing

An anaerobic laboratory-scale continuously stirred tank reactor (CSTR) of 100 L capacity was operated under mesophilic conditions as previously described (17). It was supplemented with sugar beet pulp at cautious organic loading rate (OLR) ranging from 0.5 to 2 Kg_VS m^{-3} d^{-1}, to prevent pH drop and process failure (Figure 1A). Samples were taken at regular time intervals and preserved at -80°C before the analysis. Macromolecules DNA/RNA were co-extracted with Allprep DNA/RNA Mini kit (Qiagen, Hilden, Germany), according to the manufacturer instructions. Sequencing DNA libraries were prepared using TruSeq Nano DNA kit (Illumina, FC-121-4002) using standard protocol. The libraries were prepared for 350bp average insert size RNA libraries were prepared using TruSeq stranded mRNA library preparation kit (Illumina, RS-122-2101). Libraries were sequenced with NextSeq500 (LCSB sequencing platform) using 2x150 bp read length.

2.2. Metagenomics and MAG re-construction

Following the sequencing, 361.3 megabase pairs (Mbp) of raw reads from 6 samples were quality trimmed in CLC Genomics Workbench v.9.5.2 (Qiagen), using a phred quality score of 20, minimum length of 50 and allowing no ambiguous nucleotides, resulting in close to 300.6 M quality-trimmed paired reads. Quality-trimmed reads were further co-assembled using the CLC's *de novo* assembly algorithm in a mapping mode, using automatic bubble size and word size, minimum contig length of 1000, mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 0.9, and similarity fraction of 0.95. Co-assembly resulted in 668130 contigs for a total assembly length of 2117 M nucleotides. The calculated N50 was 4334 with the longest contigs of 697049 bp. The average contig abundance was calculated as DNA-RPKMs (here equal to the number of reads mapped to the contig and normalized by the contig length and per million mappable reads). ORFs were searched and annotated using the default pipeline integrated in the IMG-Mer (18). Information about KEGGs and COGs assignment was retrieved based on the IMG-Mer annotations. To reconstruct MAGs, contigs were binned using MetaBAT2 (19) in default mode, what resulted in their separation into 3191 bins. For this study, the analysis was restricted to the 1600

MAGs harboring CAZymes coding genes, that were further assessed for completeness and contamination with CheckM using lineage-specific marker genes and default parameters (20). Taxonomic affiliation was assessed with PhyloPhlan (21). MAGs assigned to bacteria and with completeness higher than 50 % were further analysed. Using FastANI, the genome average nucleotide identity (ANI) was calculated as an indicator of the novelty of the reconstructed MAGs (22) and compared to the recently published database of 1600 MAGs reconstructed from anaerobic digestion reactors (23).

2.3. Metatranscriptomics sequencing and RNA-seq

Over 804.3 million of metatranscriptomics raw reads from 6 samples were quality trimmed in CLC Genomics Workbench v.9.5.2, using a phred quality score of 20, minimum length of 50 and allowing no ambiguous nucleotides, resulting in close to 472 M quality-trimmed reads. Quality trimmed reads were further mapped to the reconstructed metagenomics contigs with gene annotations, using the CLC "RNA-seq analysis" mode, with default parameters except for minimum similarity of 0.95 over 0.9 of the read length, both strands specificity and 1 maximum number of hits per read. The mapping results were represented as TPMs (transcripts per million), what directly resulted in normalised reads counts.

2.4. Carbohydrate-active enzymes and polysaccharide utilization loci analysis

For the purpose of this study, data analysis was restricted to contigs encoding for one or more CAZymes coding genes. The presence of putative CAZymes was identified using the CAZyme annotation web-server dbCAN (24) together with the CAZyme database (http://www.cazy.org/). Since more than one CAZyme domain can be assigned to a single CAZyme coding gene, we will refer to "CAZyme domain" whenever relevant. Annotated putative CAZymes coding genes were imported into the homology to peptide pattern (Hotpep) to further predict their enzymatic activity by being assigned to an enzyme commission (EC) class (25). Due to a high fragmentation of the data, PULs were manually identified based on the combination of at least one *sus*-like gene and one GH coding gene. Moreover, the presence of two

closely encoded CAZymes within a contig however lacking sus-like genes were refered to as clusters.

3. Results and discussion

3.1. Anaerobic reactor, a catalog of bacterial carbohydrate active enzymes

To harvest the carbohydrate hydrolytic potential and in order to acclimatize the bacterial community, the anaerobic reactor was fed with sugar beet pulp at OLR starting from 0.5 to 2 kg VS m^{-3} d^{-1} (Figure 1A). To monitor the acclimatization of the microbiome, sampling was performed at various time points, including day 1, 21, 56, 77, 118 and 132. DNA extraction followed by metagenomics analysis yielded 2 464 945 open reading frames (ORFs). Additionally, metatranscriptomic reads were mapped to the annotated ORFs (RNA-seq analysis) in order to study the gene expression profiles. Binning generated the reconstruction of 3 191 MAGs. However, only 1 300 MAGs were further assigned to a bacterial phylum and contained CAZymes. In total, 108 550 CAZymes coding genes were identified within the whole metagenome, and nearly half (49,5%) were encoded in the discussed bacterial genomes of the following origin: *Acidobacteria*, *Actinobacteria*, *Bacteroidetes*, *Chloroflexi*, *Firmicutes*, *Planctomycetes*, *Proteobacteria*, *Spirochaetes*, *Synergistetes* and *Verrucomicrobia* (Figure 1D). At the microbiome community level, the metagenomic profile of CAZymes (Figure 1A) and more specifically GHs and ECs (Supplementary Figure S1), remains similar at the different time points, highlighting the functional redundancy of the carbohydrate hydrolytic potential in AD. Indeed, it has been previously shown that AD microbiomes are able to maintain their (hydrolytic) potential despite being subject to changing environmental conditions (e.g. pH, feedstock composition) (9). However, at the gene expression level, we observe some variability over time; this variability can be an indicator of the functional plasticity of the bacterial community, i.e. the ability of the microbiome to adapt to changing environmental conditions by modulating gene expression (26).

Additionally, bacterial contribution to the relative metagenomics GHs abundance was varying (Figure 1C). At the beginning of the process, abundant GHs were essentially encoded in *Chloroflexi*, *Proteobacteria*, *Acidobacteria*, *Bacteroidetes* and *Firmicutes*

genomes. This finding supports a previous metagenomic analysis of a rice-straw enriched compost habitat where these phyla (except *Acidobacteria*) were identified as the main contributors of CAZymes abundance (27). Interestingly, after 77 days, the abundance of GHs genes, identified from *Actinobacteria* genomes, increased (Figure 1C). The abundance of *Actinobacteria* in AD has been already highlighted in previous studies, e.g. in a mesophilic anerobic digester (28) or in an anaerobic digestion sludge (29), and its contribution to the abundance of CAZymes has been shown in a compost habitat enriched with rice straw (27).

Regarding the carbohydrate hydrolytic diversity, *Chloroflexi* is the phylum encoding the largest pool of CAZymes (14 044), followed by *Bacteroidetes* (10 090), *Proteobacteria* (7 613) and *Planctomycetes* (5 707) (Figure 1D). Eventhough *Planctomycetes* genomes were enriched in CAZymes (5 707) and GHs (3 130), abundant GHs in our reactor were not assigned to this phylum. (Figures 1C,D,E). Finally, *Actinobacteria* genomes encoded for only 2 770 CAZymes and 759 GHs conserved domains. At the microbiome level, GH13 appears to be the most abundant family, followed by GH43, GH2, GH23 and GH3 (Supplementary Figure S1).

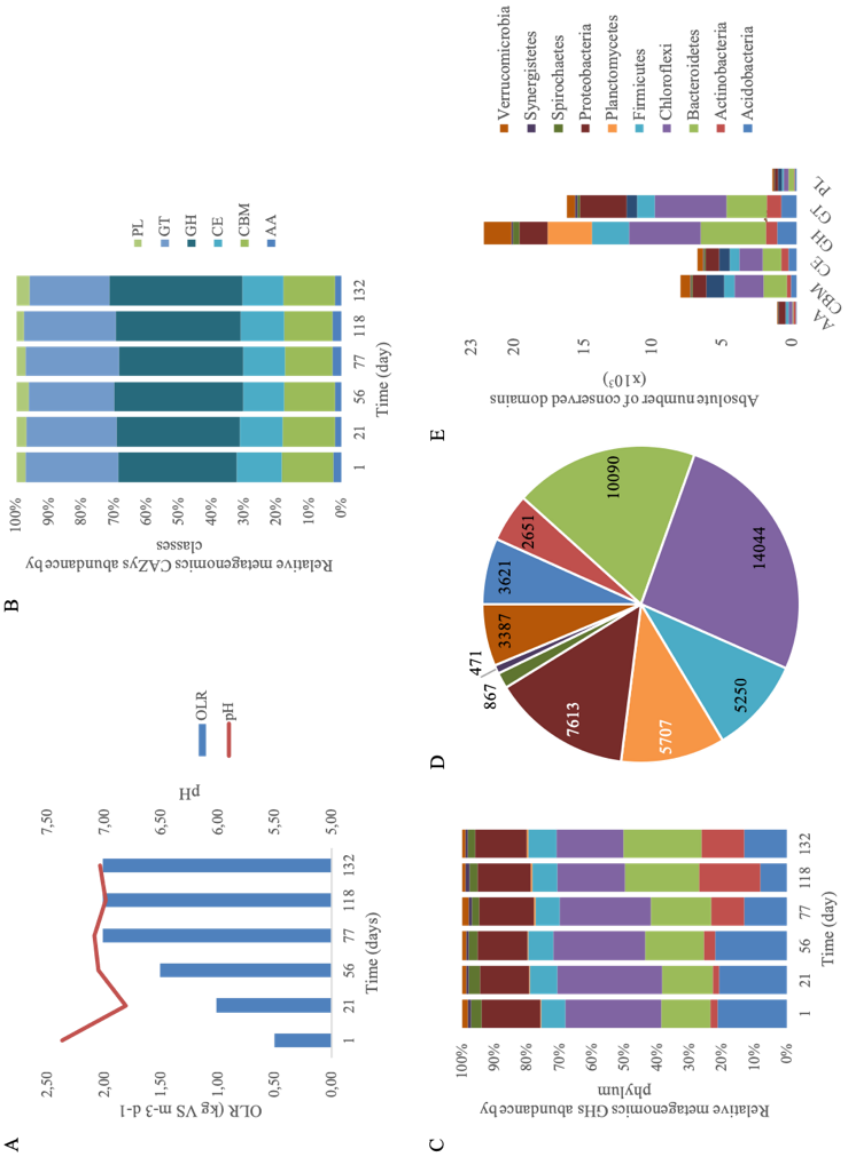


Figure 1: Carbohydrate hydrolytic potential of the microbiome over the anaerobic digestion (AD) process. **A** - Parameters applied to the AD process over time. Organic loading rate (OLR) of sugar beet pulp is represented by blue bars and expressed in kg VS m^{-3} d^{-1}. The reactor pH is represented by a red line. **B** - Relative metagenomic abundance of CAZymes coding genes coloured by CAZymes classes, over the AD process time. **C** - Relative metagenomic abundance of glycoside hydrolases (GHs) coloured by phylum, over the AD process time. **D** - Absolute number of CAZymes coding genes identified in the microbiome, coloured by phylum. **E** - Absolute number of CAZymes conserved domains identified in the microbiome, coloured by phylum. PL: Polyssacharide lyase, GT: Glycoside transferase, CE: carbohydrate esterase, CBM: carbohydrate binding module, AA: auxiliary enzymes

3.2. Glycoside hydrolases distribution as an indicator of carbohydrate hydrolytic potential

The distribution of GH coding genes was studied within the bacterial community of the investigated anaerobic digester microbiome (Figure 2), as an indicator of the carbohydrate hydrolytic potential. With 1373 identified genes, family GH13 was one of the most represented GH family (Figure 2A). Enzymes belonging to GH13 family are known to be involved in starch deconstruction (Figure 2C, (30)). In this bacterial community, starch deconstruction potential was mainly assigned to *Chloroflexi* genomes and related genes were encoding for putative α -amylases (EC 3.2.1.1). Within sugar beet pulp, hemicellulose and pectin are the two major components, while cellulose is less abundant (31). Hemicellulose deconstruction potential can be related to the presence of genes from GH2, GH43, GH3 (Figure 2C). In our study, they are the most represented families within the bacterial community, following GH13 family. Respectively, they account for 974, 931 and 868 genes (Figure 2A). The deconstruction potential of hemicellulose was mainly attributed to *Bacteroidetes* and to a lower extend to *Chloroflexi* and *Firmicutes*, and the related genes were encoding for putative β -glucosidases (EC 3.2.1.21), β-xylosidases (EC 3.2.1.37), β-mannosidases (EC 3.2.1.25) and β-galactosidases (EC 3.2.1.23) (Figure 2A/B). Similarly to hemicellulose, pectin deconstruction relies on the presence of enzymes from GH2 and GH43 families, encoding for β-galactosidases (EC 3.2.1.23) together with ARAF (EC 3.2.1.55) and arabinan endo-1,5-alpha-L-arabinosidase (EC 3.2.1.99), (Figure 2C). On top of that, *Bacteroidetes* and *Chloroflexi* also seem to have the potential of deconstructing cellulose due to the presence of putative GH5 genes in their genomes (Figure 2A). Indeed, GH5 family usually harbours endocellulases (EC 3.2.1.4) together with endomannanases (EC 3.2.1.78) (Figure 2C). Regarding less represented bacterial phyla in our anaerobic digestion microbiome, *Planctomycetes* carbohydrate hydrolytic potential is non negligible, and the second more represented family within its genome is GH5 family (involved in cellulose and hemicellulose deconstruction) (Figure 2A/C). Finally, other phyla such as *Acidobacteria*, *Actinobacteria*, *Proteobacteria*, *Spirochaetes*, *Synergistetes* and *Verrumicrobia* are less rich in CAZyme coding genes and their carbohydrate hydrolytic potential is more

oriented towards non lignocellulolytic deconstruction (GH33) as well as peptidoglycan deconstruction (GH23).

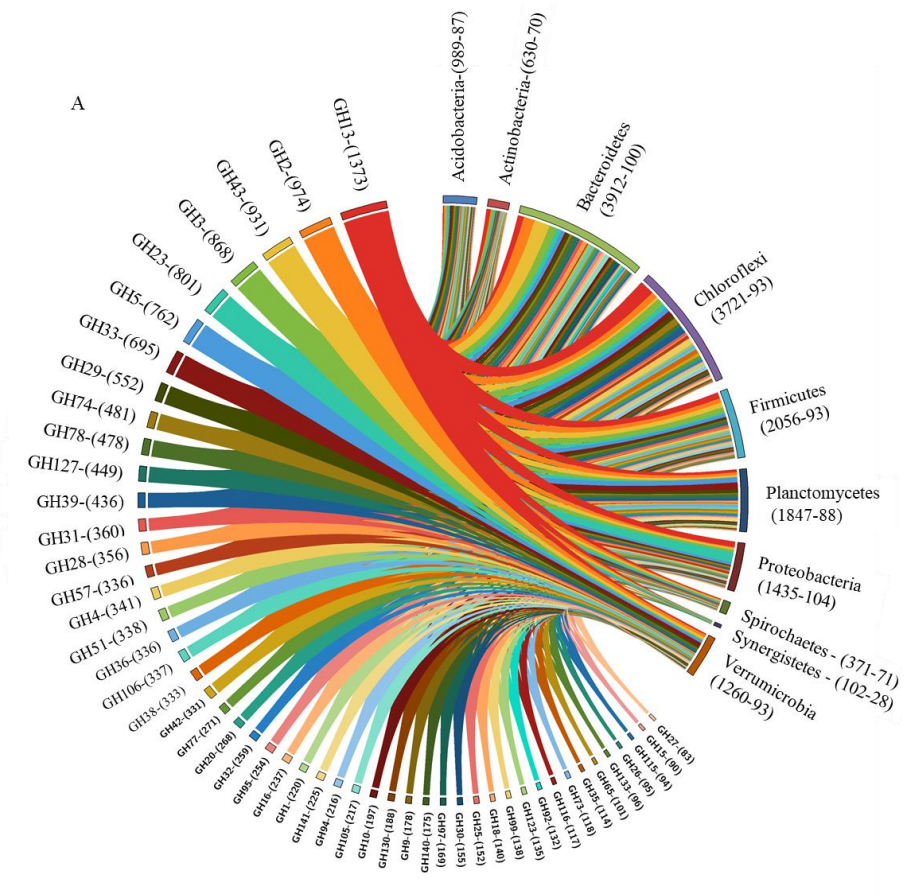
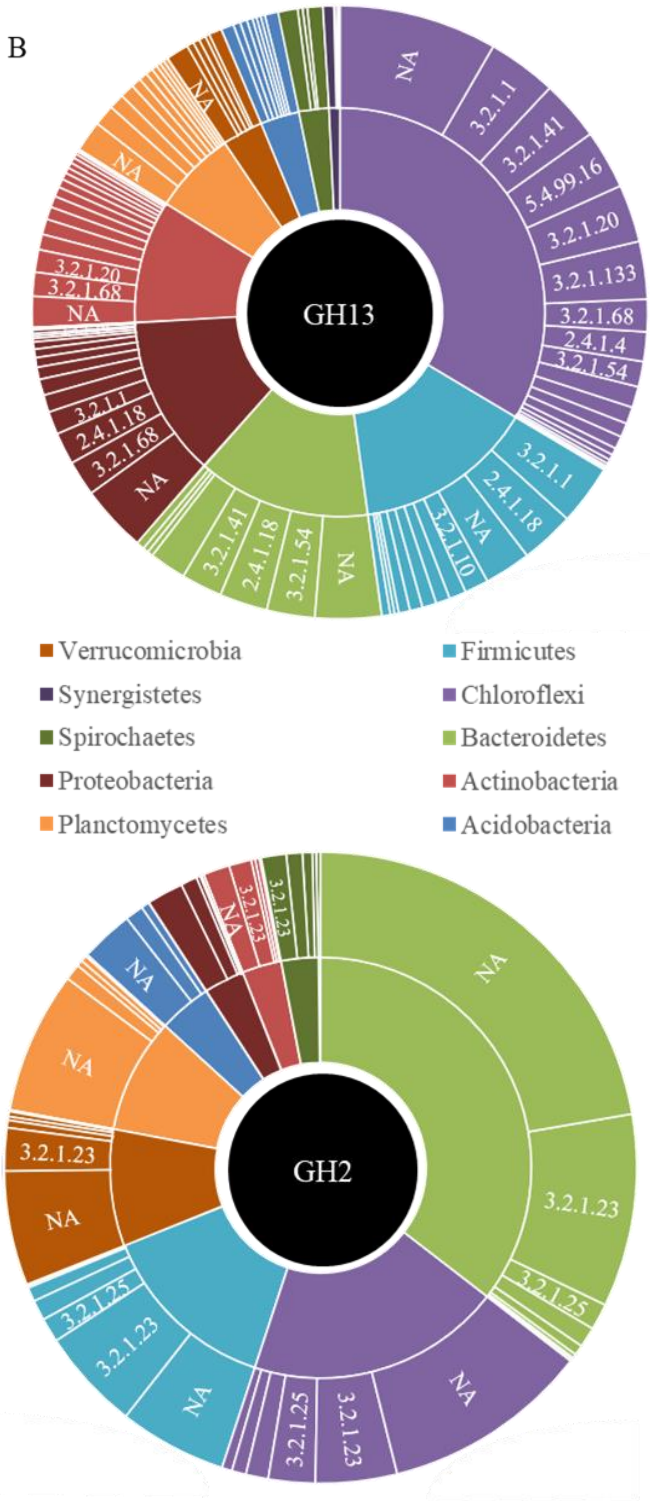
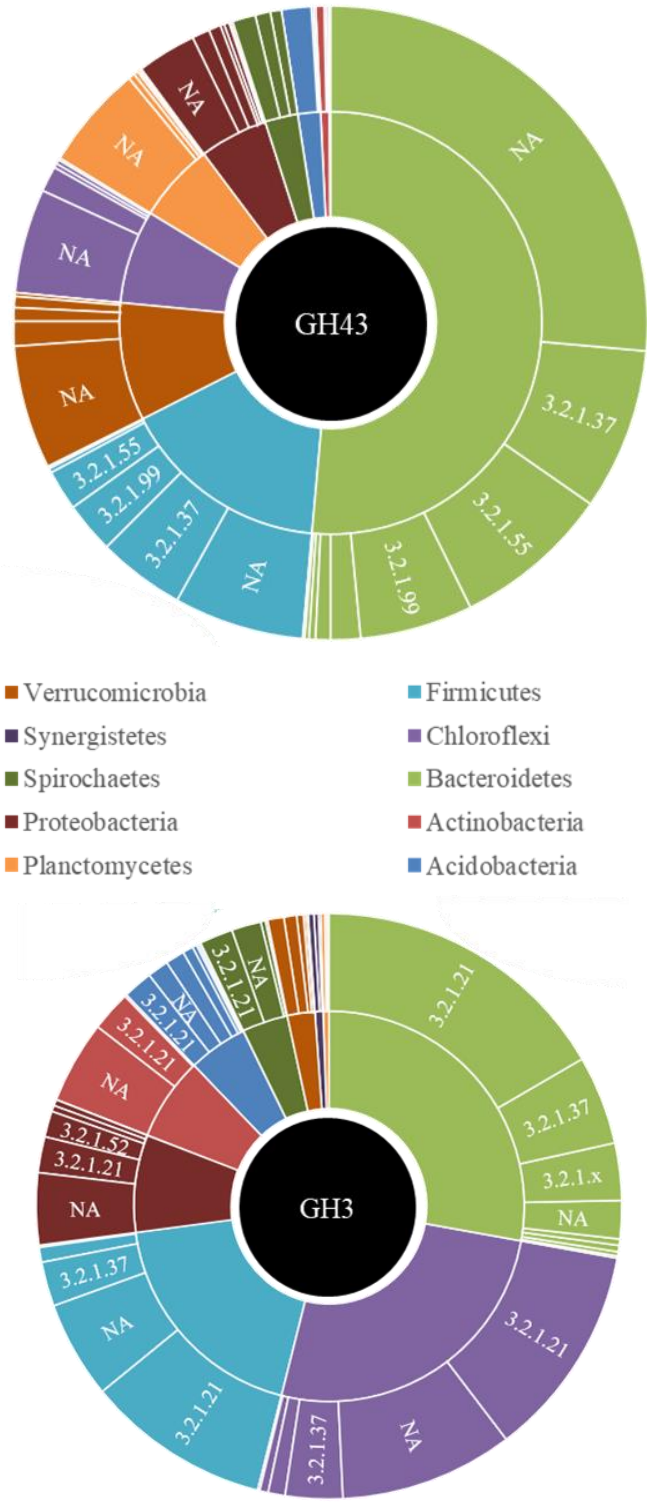


Figure 2A: Diversity of glycoside hydrolases (GHs) and their phylogenetic distribution in studied phyla present in the anaerobic digestion microbiome. GH genes distribution of the 50 more represented GH families (left) within the different studied phyla (right), the first number in the brackets refers to the number of identified GH coding genes, while the second (in phyla) refers to the CAZymes diversity within the phylum, i.e. the number of CAZymes families.


B
GH13
NA
3.2.1.1
3.2.1.41
5.4.99.16
3.2.1.20
3.2.1.133
3.2.1.68
2.4.1.4
3.2.1.54
3.2.1.1
2.4.1.18
NA
3.2.1.10
3.2.1.41
2.4.1.18
3.2.1.54
NA
3.2.1.1
2.4.1.18
3.2.1.68
NA
3.2.1.20
3.2.1.68
NA
NA
NA
Verrucomicrobia
Firmicutes
Synergistetes
Chloroflexi
Spirochaetes
Bacteroidetes
Proteobacteria
Actinobacteria
Planctomycetes
Acidobacteria
GH2
NA
3.2.1.23
3.2.1.25
NA
3.2.1.23
3.2.1.25
NA
3.2.1.23
NA
NA
NA
3.2.1.23
3.2.1.23


GH43
NA
3.2.1.37
3.2.1.55
3.2.1.99
NA
3.2.1.37
3.2.1.99
3.2.1.55
NA
NA
NA
NA
NA
Verrucomicrobia
Synergistetes
Spirochaetes
Proteobacteria
Planctomycetes
Firmicutes
Chloroflexi
Bacteroidetes
Actinobacteria
Acidobacteria
GH3
3.2.1.21
3.2.1.37
3.2.1.x
NA
3.2.1.21
NA
3.2.1.37
3.2.1.21
NA
3.2.1.52
3.2.1.21
NA
3.2.1.21
NA
3.2.1.21
NA
3.2.1.21
NA

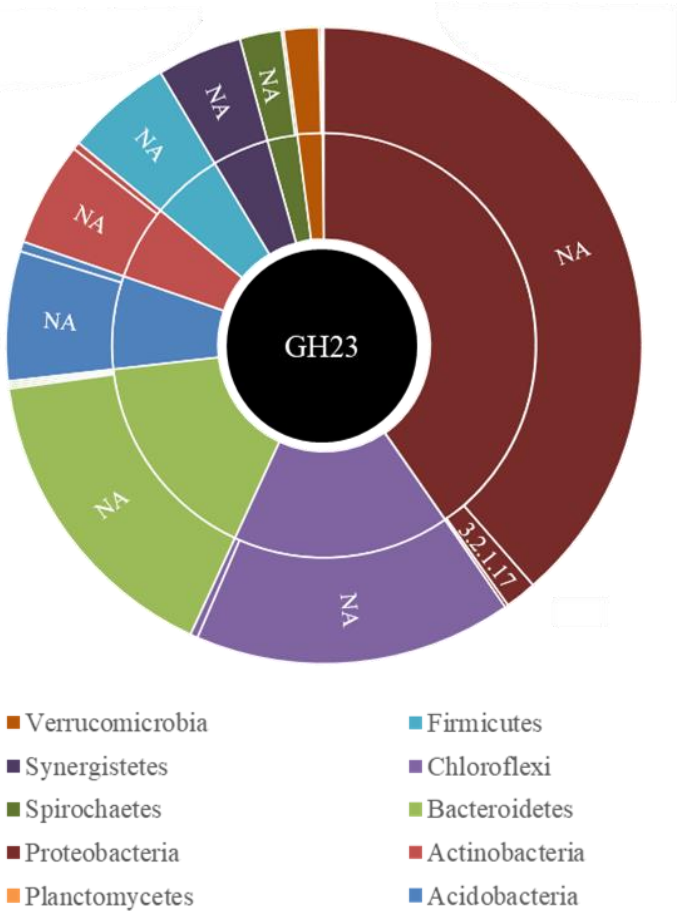


Figure 2B: Glycoside hydrolases (GHs) distribution and main activities (expressed as EC number) detected in studied phyla present in the anaerobic digestion microbiome. Distribution of genes (expressed in absolute number) and their corresponding EC assignement, identified from the five most represented GH families belonging to the studied phyla of the anaerobic digester microbiome. NA refers to Not Assigned.

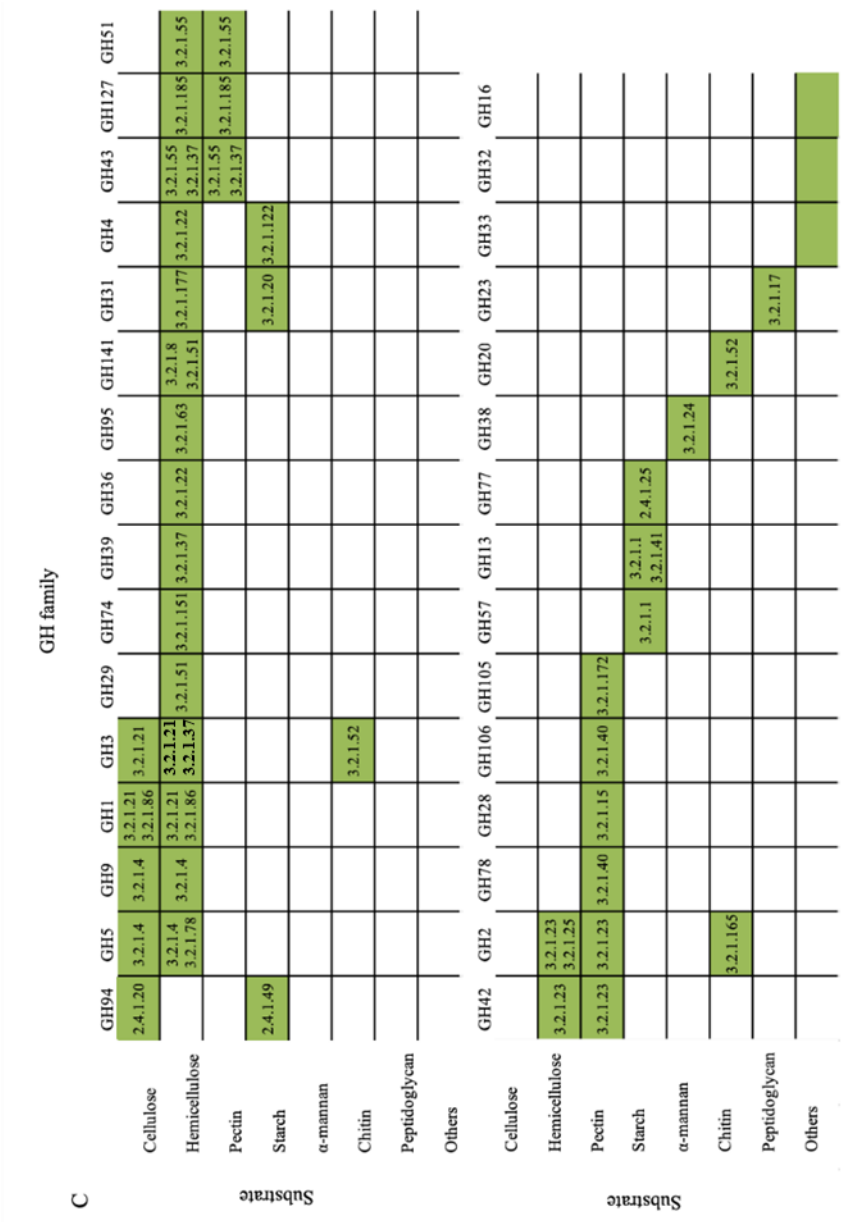


Figure 2C: Susbtrate specificity, based on the CAZymes database, of the 31 more represented glycoside hydrolase (GH) families in the anaerobic digestion microbiome. The list is not exhaustive and is based on the characterized protein available in the database.

3.3. Insight into reconstructed metagenome assembled genomes

De novo metagenome assembly resulted in the reconstruction of 198 MAGs with completeness above 50% and assigned to the different phyla (Figure 3A). Comparison of the reconstructed MAGs to a database of around 1600 MAGs from ADs, was performed by calcuting the genome average nucleotide idendity (ANI)(22). The results presented here showed the novelty of our MAGs (Supplementary Table S1). In total, 83 MAGs were newly reconstructed (calculated ANI was lower than 75%), 62 MAGs were closely related, i.e. at the genus level, to the already reconstructed MAGs from previous studies (calculated ANI was between 75% and 95%) and finally 53 MAGs were previously reconstructed (calculated ANI>95%). Therefore, this reinforce the idea that the AD microbiome involves a consortium of poorly characterised microorganisms (32–34).

Most of the MAGs reconstructed in our study represented *Proteobacteria* with 41 MAGs (21 MAGs newly reconstructed), followed by *Bacteroidetes*, *Chloroflexi* and *Firmicutes* with respectively, 14, 14 and 20 newly reconstructed MAGs, (Figure 3A and supplementary Table S1). *Verrucomicrobia* and *Planctomycetes* MAGs accounted for 9 (4 newly reconstructed) and 5 (all closely related to previously reconstructed MAGs), respectively, showing that the diversity of *Planctomycetes* in AD reactors is relatively low and the same species (based on their reconstructed genomes) are recovered from different AD locations (23). By exploring their genomes, it was noticed that they were highly rich in CAZymes and GH coding genes (Figures 3A and 3B) (between 8 and 10 % of their genes were assigned to CAZymes, resulting in the presence of between 300 and 1300 GHs in *Planctomycetes* MAGs and between 70 and 500 GHs in *Verrucomicrobia* MAGs), pointing out their high carbohydrate hydrolytic potential. Additionnally, the functional diversity of *Planctomycetes* was quite high, between 24 and 83 GH families were represented in the different MAGs (Figure 3B). Even though their genomes were highly rich in GH coding genes, there were almost not expressed during the anaerobic digestion process, e.g on average, around 20% of the identified GHs were expressed (Figure 3B). Contrariwise, *Synergistetes* and *Spirochaetes* genomes (respectively 6 and 7 MAGs identified) were poor in CAZymes and GHs but their average expression equalled around 50%.

Including all the phyla, more than 50% of the GHs were expressed on average. However, *Bacteroidetes* was identified as the phylum showing the highest GHs expression ratio. Indeed *Bacteroidetes* MAGs, were rich in GHs that on average 90% were expressed. This highlights the importance of *Bacteroidetes* in the anaerobic digestion process.

Expressed GH genes involved in the deconstruction of lignocellulosic biomass (Figure 3) were mainly identified from *Bacteroidetes* and *Chloroflexi* genomes (Figure 4). *Chloroflexi* showed a high potential for the digestion of (hemi)cellulosic biomass by expressing GH94, GH1, as well as GH39 CAZymes, which were almost not present within the *Bacteroidetes* MAGs. However, the latest showed more ability to orient its deconstruction towards pectins and hemicelluloses, especially by expressing GH43 and GH2 CAZymes. Indeed, *Bacteroidetes* carbohydrate hydrolytic potential highlights its ability to deconstruct hemicellulose and pectin (by being rich and expressing CAZymes belonging to GH43 and GH2 families), which are two major components of the sugar beet pulp (31). However, complete hydrolysis of SBP requires the presence of cellulose degrading enzymes, which were potentially expressed by *Chloroflexi* genomes. Indeed, potential cellulose degradation by *Chloroflexi* has been mentioned previously (27, 35).

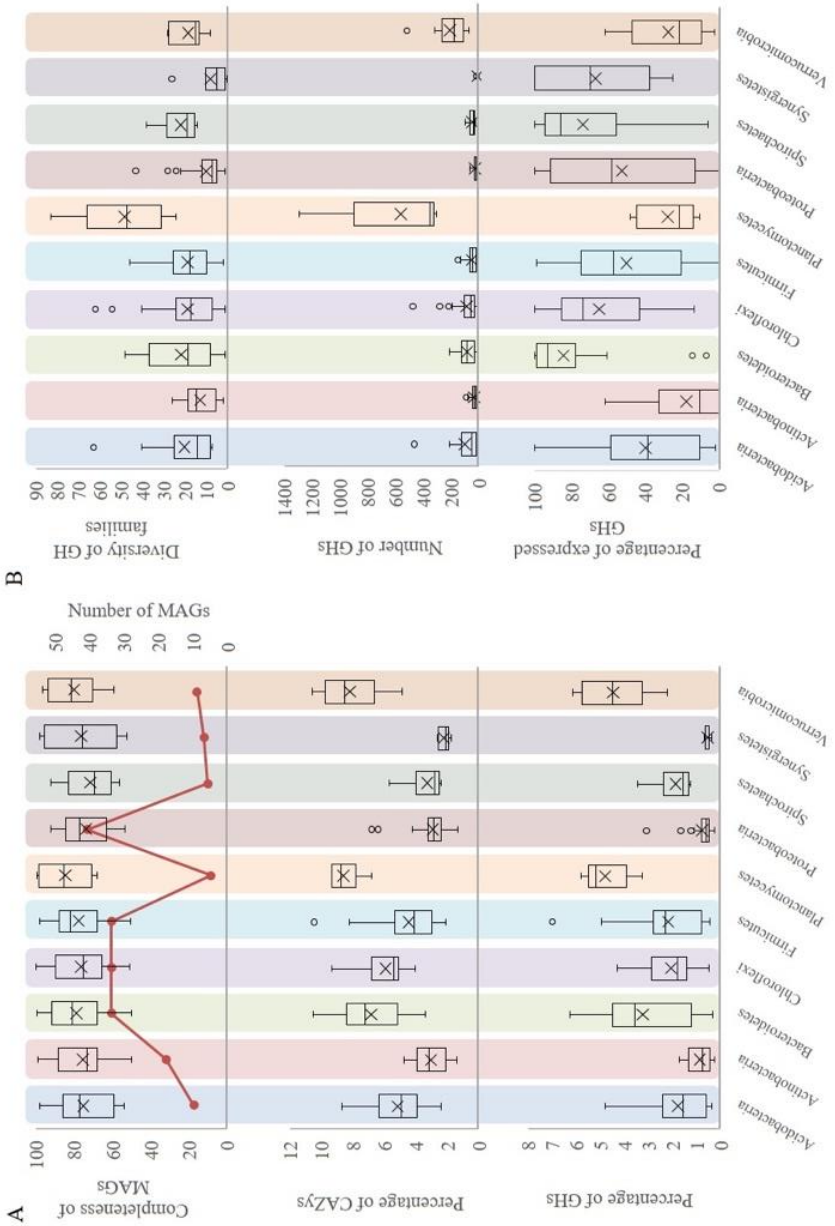
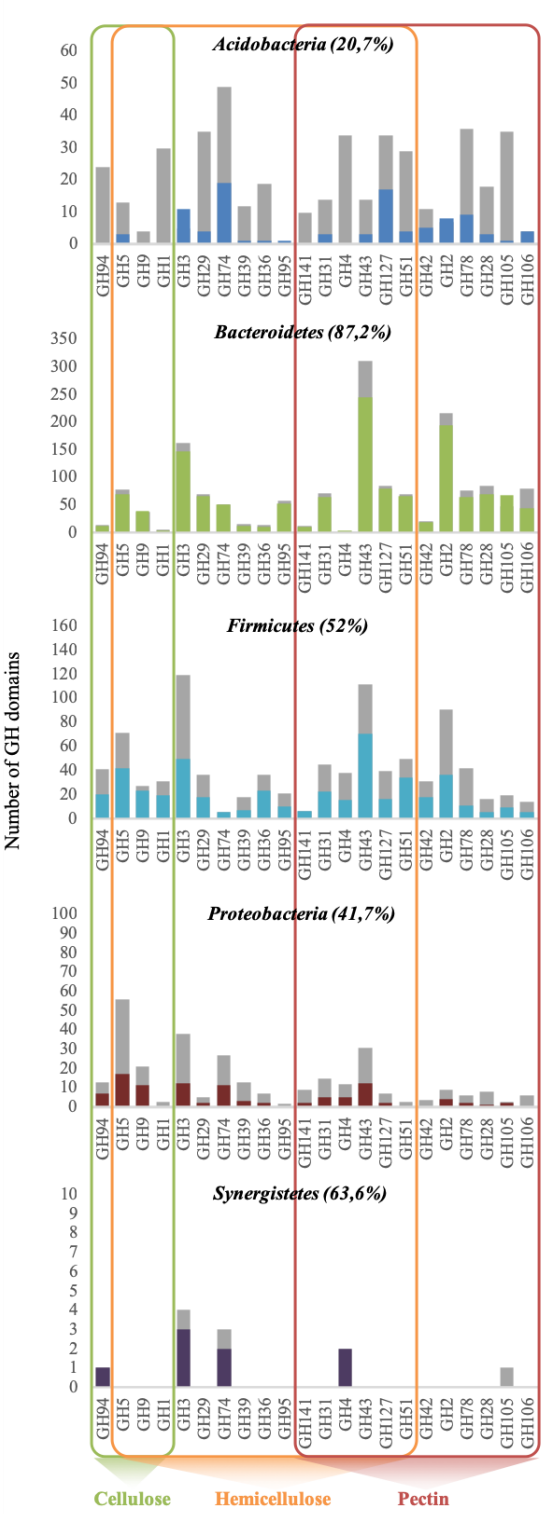


Figure 3: Box-plots of the various characterized metagenomes assembled genomes (MAGs) identified in our anaerobic digestion microbiome, resulting in 198 MAGs with completeness above 50%. **A** - Completeness of MAGs, percentage of CAZymes within MAGs and percentage of glycoside hydrolases (GHs) within MAGs are represented. The red line represents the number of MAGs assigned to each phyla. **B** - Diversity of GH families (i.e. number of GH families) in MAGs, Total number of GHs in MAGs and percentage of expressed GHs (out of total number of GHs present in the genomes) in MAGs are represented. Box-plot: median, first and third quartile, minimum and maximum


Acidobacteria (20,7%)
Bacteroidetes (87,2%)
Firmicutes (52%)
Proteobacteria (41,7%)
Synergistetes (63,6%)
Number of GH domains
GH94 GH5 GH9 GH1 GH3 GH29 GH74 GH39 GH36 GH95 GH141 GH31 GH4 GH43 GH127 GH51 GH42 GH2 GH78 GH28 GH105 GH106
Cellulose
Hemicellulose
Pectin

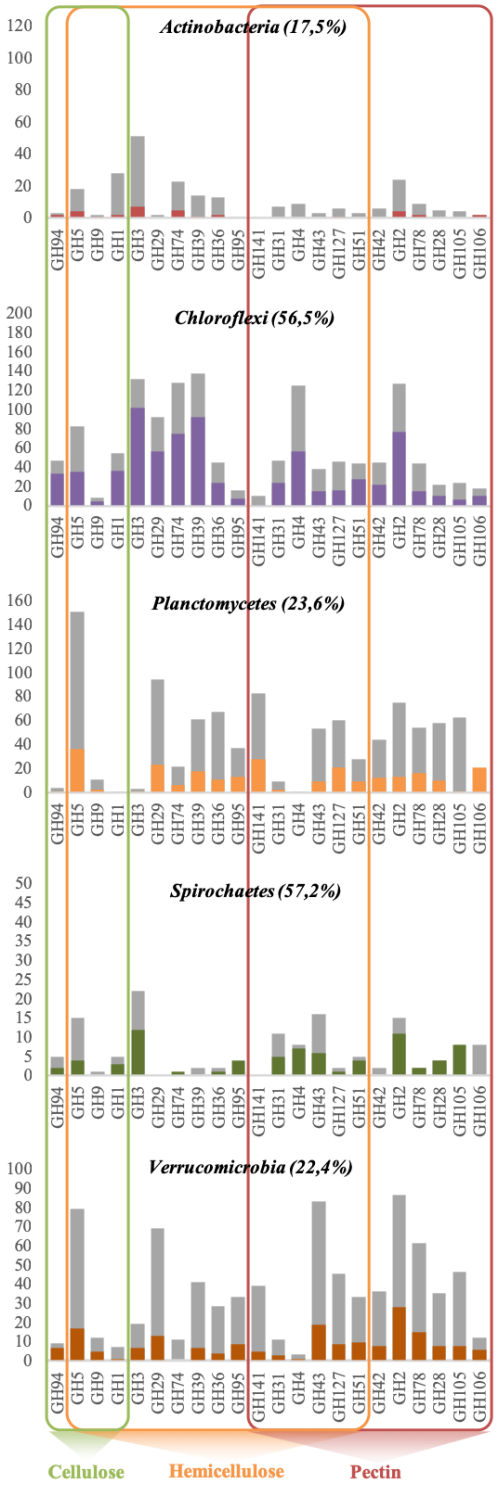

Actinobacteria (17,5%)
Chloroflexi (56,5%)
Planctomycetes (23,6%)
Spirochaetes (57,2%)
Verrucomicrobia (22,4%)
GH94 GH5 GH9 GH1 GH3 GH29 GH74 GH39 GH36 GH95 GH141 GH31 GH4 GH43 GH127 GH51 GH42 GH2 GH78 GH28 GH105 GH106
Cellulose
Hemicellulose
Pectin

Figure 4: Multi-omic analysis of the 22 main represented glycoside hydrolase (GH) families involved in lignocellulose deconstruction in the investigated anaerobic digester fed with surgar beet pulp. Grey bars represent the number of identified GH domains from metagenomics, coloured bar represent the proportion that was expressed based on metatranscriptomic analysis. The percentage of total expressed GHs (out of the number of identified GHs) per phylum is given in brackets. GH families are classified by substrate specificity (cellulose and/or hemicellulose and/or pectin).

3.4. Genome distribution of α-L-arabinofuranosidases and feruloyl esterases

As stated above, ARAFs and FAEs play a crucial role in the deconstruction of lignocellulose, by detaching cellulose and hemicellulose from lignin. Additionally, the clean isolation of ferulic acid is also an attractive industrial pathway as it is recognized as a strong antioxidant slowing down cell aging (36). Moreover, in bacterial genomes, CAZymes are either encoded alone, or form co-localized genes clusters (i.e. closely encoded CAZymes). In *Bacteroidetes*, clusters of CAZymes coding genes can be encoded together with specific transporters coding genes, i.e. Sus-like genes, known as PULs (7).

A logical justification for bioprospecting enzymes originating from the same organism and/or gene cluster, is that they are more likely to work synergistically, because they co-evolved together (37). In this sense, it is worth analyzing the distribution of specific CAZymes (they can be encoded alone, in PULs or clusters) within one single genome. Accordingly, we decided to put some efforts towards the characterization of ARAFs and FAEs, two accessory enzymes of interest. Indeed, several studies showed the impact of ARAFs on the saccharification efficiency as well as on xylose release. For example, a previous study concluded that overproduction of this enzyme increased released cellulose (probably due to the disruption of hydrogens bonds between cellulose and arabinoxylan) as well as the saccharification efficiency (38). Additionally, other researchers showed that the addition of ARAFs to enzymatic cocktails, on chemically pre-treated corn stover, increased the yield of xylose (39, 40). Similarly, several studies have established that FAEs act synergistically with endo acting enzymes (e.g. mannanases, xylanases) to deconstruct biomass (41–43), by enhancing the accessibility of (hemi)cellulases to their substrates (13, 44). Thus, for the biorefinery sector, it is essential to mine, discover and characterize novel ARAFs.

In this study, putative ARAFs and FAEs were screened based on their annotation to GH or CE CAZymes families prior being associated to their respecitive putative activity by the Hotpep analysis, i.e. EC 3.2.1.55 for ARAFs and EC 2.1.1.73 for FAEs (25). As reported within the CAZyme database (http://www.cazy.org), ARAFs were mainly identified in GH43 and GH51 families, while all known FAEs are classified in CE1 family. The same observations were made by analyzing our data. Expectedly, ARAFs were mainly encoded in *Bacteroidetes* genomes (Table 1). Additionally, almost 90% of the putative *Bacteroidetes*-isolated ARAFs were expressed. The second main source of ARAFs was identified as *Firmicutes*, however showing a lower expression ratio (62.3%). Similarly, FAEs were mainly found within *Bacteroidetes* genomes and 93.5% of them were expressed.

Because of the high diversity, together with the high expression level of ARAFs and FAEs in *Bacteroidetes* genomes, the study of the two accessory enzymes was restricted to *Bacteroidetes* genomes. Clustering of ARAFs and FAEs within *Bacteroidetes* genomes showed that only 32% and 13.8% of the expressed ARAFs and FAEs respectively, were encoded alone, while 68% and 86.2% respectively, were found either in clusters (i.e. close to at least one other CAZyme) or in PULs (Table 2). However, it should be mentioned that the number of identified ARAFs and FAEs might be underestimated. Indeed, 75.1% of the putative CAZymes from CE1 and 52.3% of the CAZymes from GH43 families could not be further assigned to an EC category. For comparison, only 15.2% of the CAZymes from GH51 could not be further assigned to an EC number.

Table 1: Distribution of the feruloyl esterases and α-L-arabinofuranosidases within the different phyla. Gene number refers to the number of genes identified at the metagenome level. Gene transcript refers to the number of gene transcripts identified at the metatranscriptome level

Phylum	Feruloyl esterases EC 3.1.1.73		α-L-arabinofuranosidases EC 3.2.1.55	
	Gene number	Gene transcript	Gene number	Gene transcript
Acidobacteria	4	2	8	3
Actinobacteria	0	0	1	0
Bacteroidetes	31	29	109	97
Chloroflexi	4	0	35	23
Firmicutes	2	2	53	33
Planctomycetes	6	1	16	5
Proteobacteria	4	2	4	0
Spirochaetes	0	0	10	4
Synergistetes	0	0	0	0
Verrucomicrobia	5	2	27	11

Table 2: Localization of the feruloyl esterases and α-L-arabinofuranosidases within the *Bacteroidetes* genomes

Genome localisation	Feruloyl esterase 3.1.1.73	α-L-arabinofuranosidase 3.2.1.55
in PULs	4	28
in cluster	21	38
isolated	4	31

Keeping in mind the objective of this study, i.e. bioprospecting accessory enzymes that can potentially enhance lignocellulose deconstruction, deeper analysis of arabinoxylan-targeting PULs was performed. The main expressed PULs identified harboring ARAFs were analysed (Figure 5 and Supplementary Figure S3). The identified (glucurono)arabinoxylan (gax) targeting PULs (gaxPUL) contains ARAFs belonging to GH43 and GH51 (Figure 5). Almost all of these gaxPUL contained at least one esterase (except PUL Ga0302357_1046105), from CE1 and CE6 families mainly, and encoding for FAEs (EC 3.1.1.73) or acetylxylan esterase (EC 3.1.1.72), respectively. Acetylxylan esterases were largely investigated in the last years due to their complementary action in xylan hydrolysis (45–48). Indeed, acetylxylan esterases break down ester link between acetic acid and xylan, increasing lignocellulose hydrolysis (49).

Additionally, the analysed PULs harboured endoxylanase (EC 3.2.1.8) activity, coming from GH10, GH43 and GH5, as well as β-xylosidase (EC 3.2.1.37), belonging to GH43 and GH3 families. In some cases, (e.g. Ga0302357_1045369, Ga0302357_1004978 and Ga0302357_1046105) PULs targeted glucuronoarabinoxylan as they encoded for putative α-glucuronosidase (EC 3.2.1.139) from GH67 CAZyme family, or xylan specific α -glucuronosidase (EC 3.2.1.131) belonging to GH115 family. Details for EC assignement are given in supplementary Table S2.

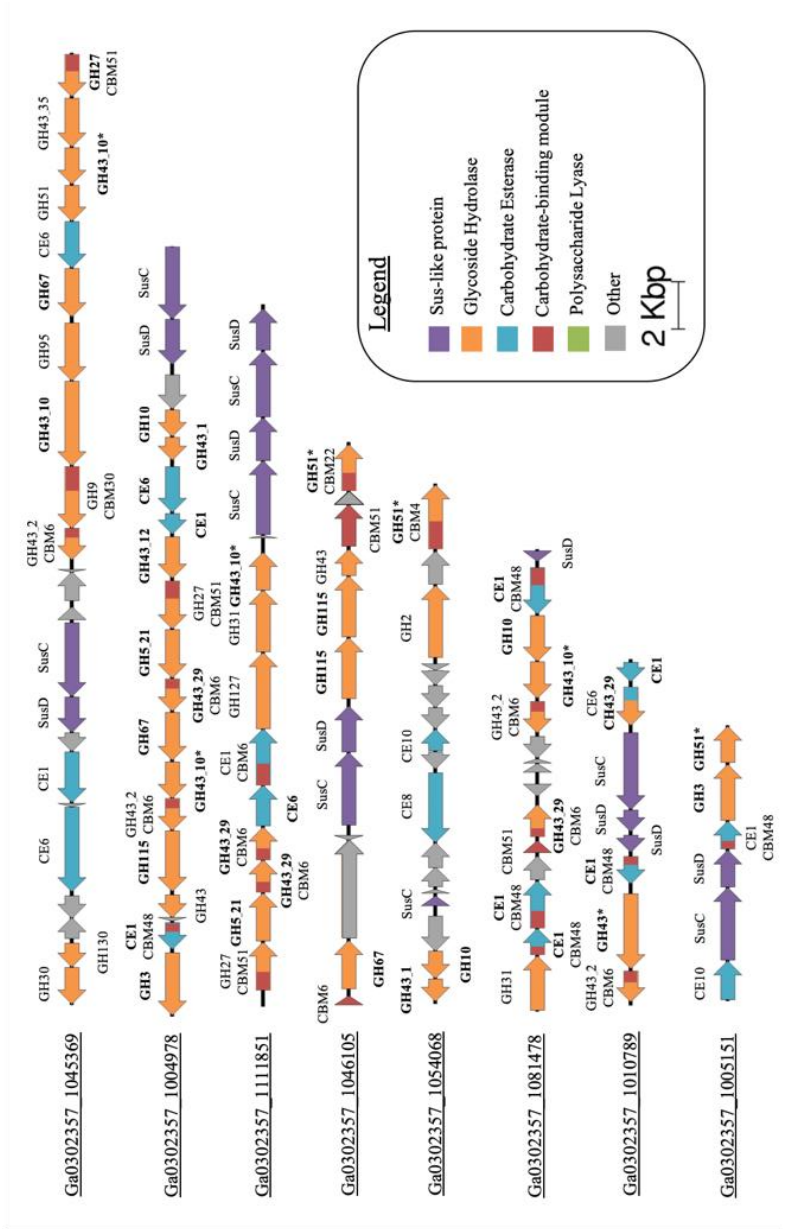


Figure 5: Putative (glucurono)arabinoxylan targeting polysaccharide utilization loci (PULs) harbouring a-L-arabinofuranosidase (EC. 3.2.1.55) activity, isolated from the microbiome of an anaerobic digester fed with sugar beet pulp. Further assignement to an EC category is represented by a bold font, and asterix identifies the putative a-L-arabinofuranosidase coding gene (further assigned to EC 3.2.1.55). Only expressed PULs are represented.

We found one putative PUL (Ga0302357_1162480, supplementary Figure S3) involved in xyloglucan degradation (the only one harboring GH29, family encoding for α-L-fucosidase (50), details for EC assignement are given in supplementary Table S3). Additionnally Ga0302357_1162480 contains putative enzymes involved in arabinoxylan (e.g. feruloyl esterase (EC 3.1.173), xylan (EC 3.2.1.37)), as well as pectin (e.g. rhamnogalacturonan α-1,2-galacturonohydrolase (EC 3.2.1.173)) deconstruction. Therefore, this PUL is probably involved in the deconstruction of a complex substrate of the plant cell wall. Few PULs were identified putatively targeting pectin (Figure S3 and Supplementary Table S3). Some of them comprise esterases belonging to many different CAZymes families (e.g. CE1, CE11, CE8, CE12). However, only few of them were assigned to an EC category (less than 28%). Thus, highlighting the low number of the currently discovered and characterized enzymes from CE family. Addtionally, putative PULs targeting pectin comprise at least one rhamnogalacturonan degrading enzyme from GH105, GH138, GH106 or GH28 CAZymes families. In some cases, PULs (e.g Ga0302357_1021651, Ga0302357_1021916 and Ga0302357_1031052) encoded for enzymes targeting pectic arabinan (i.e. endoarabinases, EC 3.2.1.99 or β-L-arabinofuranosidases, EC 3.2.1.185) from GH43 or GH127 CAZymes families.

As shown above, PULs analysis provides insight into the strategy employed by *Bacteroidetes* towards polysaccharide deconstruction, and more specifically, offers to the scientific community a catalog of enzymes for biomass hydrolysis. Additionally, such naturally-evolved enzyme cocktails open the doors to “nature-inspired formulations” of defined, minimal component enzymatic solutions specifically tailored to target the key plant polymers (e.g. arabinoxylan or pectin).

4. Conclusion

In this chapter, the carbohydrate hydrolytic potential of an anaerobic digester fed with sugar beet pulp was mined. Using MG, MAGs were reconstructed and CAZymes were identified. Using such approach, we highlighted that *Bacteroidetes*, *Chloroflexi*, *Firmicutes* and *Planctomycetes* carbohydrate hydrolytic potential was oriented towards the deconstruction of sugar beet pulp components (i.e. hemicellulose, pectin and to a lower extend cellulose). More interestingly, by an involvement of MT approach, we identified the key bacterial players of sugar beet pulp deconstruction,

namely *Bacteroidetes* and *Chloroflexi*, while *Firmicutes* and *Planctomycetes* did not take full advantage of their carbohydrate hydrolytic potential in the studied environment. Finally, due to their crucial role in lignocellulose deconstruction (by detaching lignin from hemicellulose and cellulose), ARAFs and FAEs were more deeply investigated. Genomic analyses revealed their clustering in PULs as well as in CAZymes gene clusters. These natural gene clusters might be a source of inspiration to redesign enzymatic cocktails targeting specific LCB components, what might be of interest for the biorefinery sector.

Acknowledgements

The authors greatly acknowledge Laura Lebrun and Rashi Alder for their technical support.

This work was conducted in the framework of two projects supported by the Luxembourg National Research Fund, as follows: by FNR CORE 2015 project LEGELIS (C15/SR/10404839: Linking environmental condition to lifestyle strategies and to population-level genetic heterogeneity) and by FNR CORE 2014 project OPTILYS (C14/SR/8286517: Exploring the higher termite lignocellulolytic system to optimize the conversion of biomass into energy and useful platform molecules)

References

1. D. J. Cosgrove, Growth of the plant cell wall. *Nat. Rev. Mol. Cell Biol.* **6**, 850–861 (2005).
2. P. Sarkar, E. Bosneaga, M. Auer, Plant cell walls throughout evolution: Towards a molecular understanding of their design principles. *J. Exp. Bot.* **60**, 3615–3635 (2009).
3. P. M. D. Jaramillo, H. A. R. Gomes, A. V. Monclaro, C. O. G. Silva, E. X. F. Filho, Lignocellulose-degrading enzymes: An overview of the global market. *Fungal Biomol. Sources, Appl. Recent Dev.*, 75–85 (2015).
4. J. M. Morrison, M. S. Elshahed, N. H. Youssef, Defined enzyme cocktail from the anaerobic fungus Orpinomyces sp. Strain C1A effectively releases sugars from pretreated corn stover and switchgrass. *Sci. Rep.* **6**, 1–12 (2016).
5. J. S. Van Dyk, B. I. Pletschke, A review of lignocellulose bioconversion using enzymatic hydrolysis and synergistic cooperation between enzymes—Factors affecting enzymes, conversion and synergy. *Biotechnol. Adv.* **30**, 1458–1480 (2012).
6. B. Sarrouh, Up-To-Date Insight on Industrial Enzymes Applications and Global Market. *J. Bioprocess. Biotech.* **s1** (2012).
7. J. M. Grondin, K. Tamura, G. Déjean, D. W. Abbott, H. Brumer,

Polysaccharide Utilization Loci: Fuelling microbial communities. *J. Bacteriol.* **199**, JB.00860-16 (2017).
8. E. C. Martens, N. M. Koropatkin, T. J. Smith, J. I. Gordon, Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. *J. Biol. Chem.* **284**, 24673–24677 (2009).
9. M. Bertucci, *et al.*, Carbohydrate Hydrolytic Potential and Redundancy of an Anaerobic Digestion Microbiome Exposed to Acidosis , as Uncovered by Metagenomics. *Appl. Environ. Microbiol.*, 1–16 (2019).
10. G. Banerjee, J. S. Scott-Craig, J. D. Walton, Improving enzymes for biomass conversion: A basic research perspective. *Bioenergy Res.* **3**, 82–92 (2010).
11. R. D. Hatfield, D. M. Rancour, J. M. Marita, Grass cell walls: A story of cross-linking. *Front. Plant Sci.* **7** (2017).
12. S. Lagaert, A. Pollet, C. M. Courtin, G. Volckaert, β -Xylosidases and α - L - arabinofuranosidases : Accessory enzymes for arabinoxylan degradation. *Biotechnol. Adv.* **32**, 316–332 (2014).
13. D. M. Oliveira, *et al.*, Feruloyl esterases: Biocatalysts to overcome biomass recalcitrance and for the production of bioactive compounds. *Bioresour. Technol.* **278**, 408–423 (2019).
14. A. Dilokpimol, *et al.*, Diversity of fungal feruloyl esterases : updated phylogenetic classification , properties , and industrial applications. *Biotechnol. Biofuels* **9**, 1–18 (2016).
15. L. M. F. Gottschalk, R. A. Oliveira, E. P. da S. Bon, Cellulases, xylanases, β-glucosidase and ferulic acid esterase produced by Trichoderma and Aspergillus act synergistically in the hydrolysis of sugarcane bagasse. *Biochem. Eng. J.* **51**, 72–78 (2010).
16. M. G. Tabka, I. Herpoël-Gimbert, F. Monod, M. Asther, J. C. Sigoillot, Enzymatic saccharification of wheat straw for bioethanol production by a combined cellulase xylanase and feruloyl esterase treatment. *Enzyme Microb. Technol.* **39**, 897–902 (2006).
17. X. Goux, *et al.*, Microbial community dynamics in replicate anaerobic digesters exposed sequentially to increasing organic loading rate, acidosis, and process recovery. *Biotechnol. Biofuels* **8**, 1–18 (2015).
18. V. M. Markowitz, *et al.*, IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, 115–122 (2012).
19. D. D. Kang, *et al.*, MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019** (2019).
20. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
21. N. Segata, D. Börnigen, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4** (2013).
22. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).
23. S. Campanaro, *et al.*, The anaerobic digestion microbiome: a collection of 1600 metagenome-assembled genomes shows high species diversity related to methane production. *bioRxiv*, 680553 (2019).

24. Y. Yin, *et al.*, DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, 445–451 (2012).
25. P. K. Busk, B. Pilgaard, M. J. Lezyk, A. S. Meyer, L. Lange, Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics* **18**, 1–9 (2017).
26. A. Heintz-Buschart, P. Wilmes, Human gut microbiome: function matters. *Trends Microbiol.* **26**, 563–574 (2017).
27. C. Wang, *et al.*, Metagenomic analysis of microbial consortia enriched from compost: New insights into the role of Actinobacteria in lignocellulose decomposition. *Biotechnol. Biofuels* **9**, 1–17 (2016).
28. A. M. Ziganshin, *et al.*, Bacteria and archaea involved in anaerobic digestion of distillers grains with solubles. *Appl. Microbiol. Biotechnol.* **89**, 2039–2052 (2011).
29. Y. Yang, *et al.*, Metagenomic analysis of sludge from full-scale anaerobic digesters operated in municipal wastewater treatment plants. *Appl. Microbiol. Biotechnol.* **98**, 5709–5718 (2014).
30. N. G. Graebin, *et al.*, *Immobilization of glycoside hydrolase families GH1, GH13, and GH70: State of the art and perspectives* (2016).
31. T. Rezic, *et al.*, Integrated hydrolyzation and fermentation of sugar beet pulp to bioethanol. *J. Microbiol. Biotechnol.* **23**, 1244–1252 (2013).
32. L. Treu, P. G. Kougias, S. Campanaro, I. Bassani, I. Angelidaki, Deeper insight into the structure of the anaerobic digestion microbial community; The biogas microbiome database is expanded with 157 new genomes. *Bioresour. Technol.* **216**, 260–266 (2016).
33. A. Grohmann, *et al.*, Genetic repertoires of anaerobic microbiomes driving generation of biogas. *Biotechnol. Biofuels* **11**, 1–13 (2018).
34. K. Ponni Keerthana, *et al.*, Microbiome digital signature of MCR genes – an in silico approach to study the diversity of methanogenic population in laboratory-developed and pilot-scale anaerobic digesters. *Access Microbiol.* **1** (2019).
35. U. C. Irvine, R. Paul, W. James, S. James, C. David, Phylogenetic Distribution of Potential Cellulases in Bacteria. *Appl. Environ. Microbiol.* **79**, 1545–1554 (2013).
36. M. Srinivasan, A. R. Sudheer, V. P. Menon, Ferulic acid: Therapeutic potential through its antioxidant property. *J. Clin. Biochem. Nutr.* **40**, 92–100 (2007).
37. S. Ballouz, A. R. Francis, R. Lan, M. M. Tanaka, Conditions for the evolution of gene clusters in bacterial genomes. *PLoS Comput. Biol.* **6** (2010).
38. M. Sumiyoshi, *et al.*, Increase in cellulose accumulation and improvement of saccharification by overexpression of arabinofuranosidase in rice. *PLoS One* **8** (2013).
39. D. Xin, X. Chen, P. Wen, J. Zhang, Insight into the role of α-arabinofuranosidase in biomass hydrolysis: cellulose digestibility and inhibition by xylooligomers. *Biotechnol. Biofuels* **12**, 1–11 (2019).
40. D. Gao, *et al.*, Hemicellulases and auxiliary enzymes for improved conversion of lignocellulosic biomass to monosaccharides. *Biotechnol. Biofuels* **4** (2011).
41. C. M. P. Braga, *et al.*, Addition of feruloyl esterase and xylanase produced on-site improves sugarcane bagasse hydrolysis. *Bioresour. Technol.* **170**, 316–324 (2014).

42. G. Mandalari, G. Bisignano, R. B. Lo Curto, K. W. Waldron, C. B. Faulds, Production of feruloyl esterases and xylanases by Talaromyces stipitatus and Humicola grisea var. thermoidea on industrial food processing by-products. *Bioresour. Technol.* **99**, 5130–5133 (2008).
43. Y. Xue, *et al.*, Effects of different cellulases on the release of phenolic acids from rice straw during saccharification. *Bioresour. Technol.* **234**, 208–216 (2017).
44. P. Debeire, P. Khoune, J. M. Jeltsch, V. Phalip, Product patterns of a feruloyl esterase from Aspergillus nidulans on large feruloyl-arabino-xylo-oligosaccharides from wheat bran. *Bioresour. Technol.* **119**, 425–428 (2012).
45. F. A. Adesioye, T. P. Makhalanyane, P. Biely, D. A. Cowan, Phylogeny, classification and metagenomic bioprospecting of microbial acetyl xylan esterases. *Enzyme Microb. Technol.* **93**–**94**, 79–91 (2016).
46. P. Biely, B. Westereng, V. Puchart, P. de Maayer, D. A. Cowan, Recent Progress in Understranding the Mode of Action of Acetylxylan Esterases. *J. Appl. Glycosci.* **61**, 35–44 (2014).
47. P. M. A. Pawar, *et al.*, Expression of fungal acetyl xylan esterase in Arabidopsis thaliana improves saccharification of stem lignocellulose. *Plant Biotechnol. J.* **14**, 387–397 (2016).
48. H. Wu, *et al.*, Heterologous Expression of a New Acetyl Xylan Esterase from Aspergillus niger BE-2 and its Synergistic Action with Xylan-Degrading Enzymes in the Hydrolysis of Bamboo Biomass. *BioResources* **12**, 434–447 (2017).
49. J. Zhang, M. Siika-Aho, M. Tenkanen, L. Viikari, The role of acetyl xylan esterase in the solubilization of xylan and enzymatic hydrolysis of wheat straw and giant reed. *Biotechnol. Biofuels* **4** (2011).
50. J. Van Den Brink, R. P. De Vries, Fungal enzyme sets for plant polysaccharide degradation. *Appl. Microbiol. Biotechnol.* **91**, 1477–1492 (2011).

Chapter 4

Integrative omics analysis of the termite gut system adaptation to miscanthus diet identifies lignocellulose degradation enzymes

Published as:

Calusinska M., Marynowska M., Bertucci M., Untereiner B., Klimek D., Goux X., Sillam-Dussès D., Gawron P., Halder R., Wilmes P., Ferrer P., Gerin P., Roisin Y. and Delfosse P. (2020). Integrative omics analysis of the termite gut system adaptation to miscanthus diet identifies lignocellulose degradation enzymes. *Communications Biology Volume 3, Issue 1*

Personal contribution:

- Identification and characterization of clusters harboring carbohydrate active enzymes
- Gene selection, experimental design and laboratory work (recombinant protein production, purification, enzymatic assays, etc.), with the help of co-authors.
- Participation to the writting

In the previous chapters (Chapters 2 and 3), I applied metagenomics and metatranscriptomics to analyse an anaerobic digestion microbiome fed with sugar beet pulp, in order to: (1) define the composition of the anaerobic digester bacterial community (2) characterize the carbohydrate hydrolytic potential of the studied microbiome (3) identify key bacterial players involved in LCB deconstruction and (4) suggest new activities to be included in nature-inspired enzymatic cocktails for the biorefinery sector (including α-L-arabinofuranosidases and feruloyl esterases, identified as key enzymes involved in lignin detachment).

To investigate a more recalcitrant substrate (rich in lignin) and a different digestion microbiome, in the present chapter (Chapter 4), I applied metagenomics and metatranscriptomics to the termite's gut system at the holobiont level (gut microbiome and host) in order to assess its dynamic adaption towards an imposed dry miscanthus diet, in order to: (1) characterize the enzymatic arsenal available in the termite gut system to deconstruct LCB (2) identify clusters of new CAZymes targeting different lignocellulose fractions, (3) and heterogolously produce and biochemically characterize potentially interesting enzymes for the biorefinery sector.

Abstract

Miscanthus *(Miscanthus x giganteus* sp.) biomass could satisfy future biorefinery value chains. However, its use is largely untapped due to high recalcitrance to deconstuction. Termite, comprising its gut microbiome, is considered the most efficient lignocellulose degrading system in nature.

Here, we investigate at holobiont level the dynamic adaptation of *Cortaritermes* sp. to imposed miscanthus diet, with a long-term objective of overcoming lignocellulose recalcitrance. We use an integrative omics approach combined with enzymatic characterisation of carbohydrate active enzymes from termite gut *Fibrobacteres* and *Spirochaetae*. Modified gene expression profiles of gut bacteria suggest a shift towards the utilisation of cellulose and arabinoxylan, two main components of miscanthus lignocellulose. Low identity of reconstructed bacterial genomes to closely related species supports the hypothesis of a strong phylogenetic relationship between host and its gut microbiome.

This study provides a framework for better understanding the complex lignocellulose degradation by the higher termite gut system and paves a road towards its future bioprospection.

Keywords: CAZymes, lignocellulose, metagenomics, (meta)transcriptomics, miscanthus, termite gut system

1. Introduction

In a world of finite biological resources, the agenda of the UN's Sustainable Development Goals challenges scientific community to develop transformative technologies that would enable the replacement of petroleum-based raw materials and energy with renewable bio-based feedstock. Plant biomass (lignocellulose), being the most abundant and renewable natural resource (1). Miscanthus is a rhizomatous grass and owing to its adaptability to various environmental conditions, it shows high potential for sustainable production of lignocellulose over large geographical range (2). Considering its important agronomic advantages (e.g. high biomass yield per hectare, reduced soil erosion and low fertilizer and pesticide requirements), it is suitable for different biorefinery value chains, including bioethanol, biogas, food additives, ingredients for cosmetics, biopharmaceuticals, bioplastics, biomaterials, organic fertilizers and animal feed (3). Yet, due to the high recalcitrance (resistance of the cell wall components to enzymatic hydrolysis), its use is largely untapped (4).

In living organisms, enzymatic hydrolysis of lignocellulose is mainly driven by carbohydrate active enzymes (CAZymes (5)). Glycoside hydrolases (GHs) are the primary enzymes that cleave glycosidic linkages. Often, they are assisted by carbohydrate esterases (CEs), polysaccharide lyases (PLs) and other auxiliary enzymes (AAs). With its unique consortium of microorganisms, the termite gut is considered as the most efficient lignocellulose degrading system in nature (6). Complete loss of gut cellulolytic flagellates in all evolutionary higher termites and acquisition of novel symbiotic bacteria led to improved lignocellulolytic strategies. It allowed for diet diversification from mainly wood-restricted to e.g. dry grass and other plant litter, herbivore dung and soil organic matter at different stages of humification (7). Until now, most research has focused on endogenous endoglucanases of termites and cellulases from termite gut flagellates (8). CAZymes from higher termite gut bacteria have only recently started receiving scientific attention (9).

Here, we investigated the higher termite gut system of *Cortaritermes* sp. (*Nasutitermitinae* subfamily) from French Guiana savannah. Using 16S rRNA gene amplicon profiling of termite gut bacteria, we investigated the adaptation of two termite colonies to miscanthus diet under laboratory conditions. Through the *de novo* metagenomic (MG) and metatranscriptomic (MT) reconstructions, we assessed the

distribution of activities within gut community, and we linked it to different bacteria, two main players being *Spirochaetae* and *Fibrobacteres*. Further analysis of gene expression profiles proved bacterial functional plasticity (adaptation to changing environmental conditions through differential genes expression), and highlighted the abundance of gene transcripts involved in carbohydrate metabolism and transport. Analysis of the reconstructed community metagenome evidenced CAZyme clusters targeting two main components of miscanthus biomass, namely cellulose and (arabino)xylan. Most of these clusters were allocated to the reconstructed metagenome assembled genomes (MAGs) of *Fibrobacteres* and *Spirochaetae* origin. The *de novo* reconstruction of the host epithelial gut transcriptome evidenced its contribution to lignocellulose degradation, and its adaptation to miscanthus diet. Based on the characterisation of purified bacterial CAZymes, we verified the *in silico* predicted activities for many backbone-targeting (e.g. endocellulases and xylanases) and debranching enzymes (e.g. arabinofuranosidases). To finish, we discussed our findings in the context of enzymes application in the developing biorefinery sector.

2. Material and methods

2.1. Nest origin, laboratory maintenance and sampling

Initially, three colonies of grass-feeding higher termites from *Nasutitermitinae* (nests: LM1, LM2 and LM3) were identified in January 2017 in tropical savannah in French Guiana, in proximity to Sinnamary town (radius of 5 km to GPS: N 05°24.195' W 053°07.664'). Termite nests were transported to the laboratory where colonies were maintained in separate glass containers at 26 °C, 12 h light and 12 h dark, and 90 % humidity conditions. Termite colonies were fed with dried miscanthus grass winter harvest rich in recalcitrant lignocellulose (10), for a period of up to nine months. Colony LM1_2 died after few months and was excluded from extended analysis. Mature worker-caste individuals were sampled in regular monthly time intervals before (sample taken before miscanthus diet corresponds to "wild-microbiome") and following miscanthus diet (samples correspond to "miscanthus-adapted microbiome"). Termite specimens were cold immobilized, surface-cleaned with 80 % ethanol and 1 x PBS and decapitated. Whole guts (here relative to the midgut and

hindgut compartments) were dissected (n ≈ 30 per replicate, minimum three replicates per sample) and preserved directly in liquid nitrogen. Additionally, for a sample selected for metagenomics analysis (LM1 time point 8 months; LM1_8) the hindgut luminal fluid was collected as previously described (11). Samples were stored at -80 °C until further processing. Termite species were identified by morphology and by sequencing of the partial COII marker gene, as described before (12). The nucleotide sequences are available in GenBank under accession numbers MN803317-19.

2.2. Extraction of nucleic acids

DNA and RNA were co-extracted from all samples using the AllPrep PowerViral DNA/RNA Kit (Qiagen) following the manufacturer's protocol. To assure the proper disruption of bacterial cell wall and termite gut epithelium cells, mechanical bead-beating step with 0.1 mm glass beads at 20 Hz for 2 min was introduced to complement the chemical lysis. The eluents were divided in two parts. First part was treated with one µL of 10 µg/mL RNase A (Sigma) for 30 min at room temperature. The second part was treated with TURBO DNA-free kit (Invitrogen) according to manufacturer's protocol. The resulting pure DNA and RNA fractions were quality assessed using agarose gel electrophoresis and Bioanalyser RNA 6000 Pico Kit (Agilent). Nucleic acid concentration was quantified using Qubit dsDNA HS Assay and Qubit RNA HS Assay Kit (Invitrogen). DNA and RNA were stored at -20 °C and -80 °C, respectively.

2.3. 16S rRNA gene amplicon high-throughput sequencing and data analysis

The bacterial 16S rRNA gene amplicon libraries were prepared using Illumina compatible approach as previously described (13). Briefly, modified universal primers S-D-Bact-0909-a-S-18 (ACTCAAAKGAATWGACGG) and S-*-Univ-*-1392-a-A-15 (ACGGGCGGTGTGTRC, (14)), and Nextera XT Index Kit V2 (Illumina) were used along with Q5 Hot Start High-Fidelity 2x Master Mix (New England Biolabs) in a two-step polymerase-chain reaction (PCR). In the first step selective amplification of the 484 bp long fragments of bacterial 16S rRNA gene V6–V8 region was performed. llumina-compatible adapters and barcodes were attached in the second step. Purified and equimolarly pooled libraries were sequenced along with PhiX

control (Illumina) using MiSeq Reagent Kit V3-600 on in-house Illumina MiSeq Platform. Usearch v.7.0.1090_win64 software (15) was used for quality trimming, chimera check, singletons removal and assignment of the obtained sequences to OTUs at 97 % similarity level. Taxonomic affiliation of the resulting OTUs was performed with SILVA database v.128 (16). Raw sequences are available in the Sequence Read Archive (SRA) database under project number PRJNA587606. Resulting OTUs were deposited in GenBank under project numbers PRJNA586754 and PRJNA434195. Downstream analyses were performed with mothur (17) and R environment (18). Bacterial richness and diversity were calculated using respectively sobs and invsimpson indices. The dissimilarity of bacterial community structures was calculated using Bray-Curtis index. OTUs differentially abundant between the wild- and miscanthus-adapted microbiomes were assessed using the LEfSe approach (19).

2.4. De *novo* metagenomics and data analysis

Sample LM1_8 was selected for metagenomic sequencing in order to reconstruct genomes/ larger genomic fragments of the dominant bacteria in the miscanthus-adapted microbiome. Prokaryotic DNA was enriched from the total hindgut DNA extract using NEBNext Microbiome DNA Enrichment Kit (NewEngland BioLabs). Following sequencing, over 170 Mbp raw reads were quality trimmed in CLC Genomics Workbench v.9.5.2 (Qiagen), using a phred quality score of 20, minimum length of 50 and allowing no ambiguous nucleotides, resulting in close to 154 million quality-trimmed paired reads. Raw sequencing reads are available in the SRA database under project number PRJNA587423. Quality-trimmed reads were assembled using the CLC's *de novo* assembly algorithm in a mapping mode, using automatic bubble size and word size, minimum contig length of 1000, mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 0.9, and similarity fraction of 0.95. The average contig abundance was calculated as DNA-RPKMs (reads per kilo base per million mapped reads). This type of normalization allows for comparing contigs (genomic fragments) coverage (abundance) values, as it corrects differences in both sample sequencing depth and contig length. ORFs were searched and annotated using the default pipeline integrated in the IMG/MER (20). Information about KEGGs and COGs assignment was retrieved based on the IMG/MER annotations. Metabolic pathways/modules were reconstructed using the tool

integrated in the online version of the KEGG database (21). Initially, contigs were binned using myCC (22) what resulted in their separation into 35 phylum-level bins of relatively high contamination. This approach was undertaken to correctly assign phylum-level taxonomy to the resulting contigs. Subsequently, the bin refinement module integrated in MetaWRAP was used to fine-tune the resulting bins (MAGs) to the species/strain levels (23). The completeness and contamination of the generated MAGs were assessed with checkM (24). Taxonomic affiliation was assessed with PhyloPhlan (25). Similarity to the previously reconstructed MAGs was verified with the FastANI (26). MAGs abundance within the reconstructed metagenome was calculated as average of contigs metagenomic abundance (DNA-RPKMs, see above) assigned to a specific MAG. Given the relatively high bacterial diversity in the termite gut, only 54.6 % of the resulting MG sequencing reads could map back to the reconstructed MG contigs, potentially mitigating the rate of functional gene discovery if solely relaying on the *de novo* MG reconstruction. Phylogenetic analyses were performed on MAFFT-aligned protein sequences (27) using MEGA X (28).

2.5. De novo metatranscriptomics, host transcriptomics and data analysis

For three selected samples (LM1-1, LM1-2 and LM1-8) the (meta)transcriptomic analysis was performed using an optimised approach described earlier (11). Ribo-Zero Gold rRNA Removal Kit "Epidemiology" (Illumina) was used to enrich the sample for prokaryotic and eukaryotic mRNA. Enriched mRNA was purified using Agencourt RNAClean XP Kit and analyzed with Bioanalyser RNA 6000 Pico Kit (Agilent). In continuation, SMARTer Stranded RNA-Seq Kit (Clontech) was used according to the manufacturer's instructions to prepare sequencing libraries. Final libraries were quantified with High Sensitivity DNA Kit (Agilent) and KAPA SYBR FAST Universal qPCR Kit. Libraries were pair-end sequenced at the Luxembourg Centre for Systems Biomedicine (University of Luxembourg) using Illumina NextSeq 500/550 High Output v2-300 Kit. Raw sequencing reads are available in the SRA database under the project number PRJNA587406. Over 285 million raw reads were quality trimmed in CLC Genomics Workbench v.9.5.2, using a phred quality score of 20, minimum length of 50 and allowing no ambiguous nucleotides, resulting in close to 214 million quality-trimmed reads. Contaminating rRNA reads were removed using

SortMeRNA 2.0 software (29). The resulting non-rRNA reads were used to perform *de novo* (meta)transcriptome co-assembly using the CLC assembly algorithm in mapping mode with default parameters, except for minimum contig length of 200, length fraction of 0.90 and similarity fraction 0.95. As a result, nearly 2 million contigs were assembled. Obtained contigs were further submitted to IMG/MER for taxonomic and functional annotation (20). Following the taxonomic assignment, 759,451 transcripts of putative prokaryotic origin were selected for further analysis. Initial IMG/MER taxonomic annotation resulted in over-representation of transcripts of putative *Firmicutes* origin (Supplementary Fig. 4). As nearly no *Firmicutes* OTUs were detected using the 16S rRNA gene amplicon sequencing, transcripts were re-annotated based on the *de novo* assembled metagenome and contig binnig, resulting in re-classification of virtually all *Firmicutes*-assigned contigs to *Fibrobacteres* and *Spirochaetae*. To complement the study and to characterise potential contribution of the termite host to miscanthus digestion, transcripts of eukaryotic origin and taxonomically assigned to Insecta (based on the IMG/MER annotation) were further evaluated for the completeness of the *de novo* reconstructed transcriptome with the BUSCO pipeline (30) and using the Eukaryota database (odb9). There were only two duplicated genes out of the total of 303 searched groups, suggesting that the level of possible contamination of non-Insecta origin was very low (below 0.7 %).

For both the *de novo* assembled metatranscriptome and termite transcriptome, in order to determine the relative abundances of transcripts across studied samples, sequencing reads were mapped back to the annotated transcript sets, using the CLC "RNA-seq analysis" mode, with default parameters except for minimum similarity of 0.95 over 0.9 of the read length, both strands specificity and 1 maximum number of hits per read. The mapping results were represented as TPMs (transcripts per million), what directly resulted in normalised reads counts.

2.6. Identification of carbohydrate active enzyme genes and enzyme characterisation

Genes (metagenomics) and gene transcripts (metatranscriptomics) encoding for bacterial carbohydrate active enzymes were detected using dbCAN (dbCAN-fam-HMMs.txt.v6; (31)) and CAZy database (32). Using the threshold of e-value of < 1e-18 and coverage > 0.35 recommended for prokaryotic CAZymes resulted in removal

of a high number of putative CAZymes, therefore additional manual curation was performed to maximise the number of entries retained for further analysis. Additionally, gene transcripts outliers (very partially reconstructed gene fragments with average MT expression significantly exceeding the average expression of other genes assigned to the same group) were manually identified and removed as they were considered as chimeric (additional Blast search was launch in each case). Homology to peptide pattern (Hotpep) was used to assign an EC class to the identified CAZymes (33). Sub-cellular localisation of CAZymes was predicted using BUSCA web (34). To link the degradation of the different lignocellulose fractions and subsequent sugar utilisation, we looked for the presence of suitable sugar transporters and also specific sugar isomerases and kinases. Eukaryotic CAZymes, including for other sequenced termite genomes were further searched with dbCAN2 and using dbCAN-fam-HMMs.txt.v8, with new AA families included.

Genes encoding for CAZymes of interest, selected based on their predicted activities and their expression profiles, were further PCR amplified (Veriti™ 96 wells Thermal cycler, Applied Biosystems, Foster City, USA) and the resulting PCR products were purified using a PCR purification kit (Qiagen, Hilden, Germany). If any signal peptide was predicted (using LipoP version 1.0, http://www.cbs.dtu.dk/services/LipoP/, (35)), it was removed before cloning to enhance cytoplasmic protein production. Purified PCR products were cloned into the pET52b(+) plasmid and expressed in E. *coli* Rosetta (DE3) strain (Millipore Corporation, Billerica, MA, USA), as previously described (36). Cells were harvested by centrifugation (5,000 x g, 4°C, 15 min) and re-dissolved in a lysis buffer (50 mM NaH2PO4, 10 mM imidazole, 300 mM NaCl, pH8). Proteins were released by sonication, cell debris were removed by centrifugation (16,000 x g, 4°C, 15 min) and subsequent filtration step (13-mm syringe filter, 0.2-µm-pore-size). Affinity tag purification was achieved using a histidine tag located at the C terminus of a recombinant protein. NGC™ Medium-Pressure Liquid Chromatography system (Bio-Rad) equipped with a HitrapTM column of 5 mL (Bio-Rad, Hercules, CA, USA) was used to purify produced proteins. A constant flow rate of 5mL/min was applied. Initially, column was equilibrated with six column volumes (CV) of lysis buffer. Following equilibration, 290 mL of sample was injected and washed with three CVs of mixed buffer (97 % of lysis buffer and 3 % of elution buffer, the later composed of 50 mM NaH_2PO_4, 250 mM imidazole, 300

mM NaCl, pH8). First step of elution was achieved using ten CVs of a linear gradient of mixed buffer (from 3 % of elution buffer plus 97 % of lysis buffer to 50 % of elution buffer plus 50 % of lysis buffer), followed by four CVs of 100 % of elution buffer and finally one CV of 100 % lysis buffer to detach the remaining protein. Fractions of two mL were collected during the washing and elution steps, and were analysed on SDS-PAGE. The release of 4-nitrophenol (PNP assay) and/or reducing sugar (RS assay) was used to determine the activity of recombinant proteins. Briefly, 50 µL of a purified protein solution was incubated with 50 µL of substrate (respectively, 4-nitrophenol derivatives were used for PNP assay and polysaccharides for RS assay) and 100 µL (PNP assay) or 25 µL (RS assay) of citrate phosphate buffer (pH7 0.1M citric acid, 0.2M dibasic sodium phosphate). The targeted substrates included carboxymethylcellulose (CMC), arabinoxylan, galactomannan, glucomannan and xylan. Enzymatic reaction was carried out at 37 °C during one hour (PNP assay) or 30 min (RS assay). The rate of release of 4-nitrophenol was instantly monitored at 405 nm using SPECORD 250 PLUS (Analytic Jena). The release of reducing sugars was determined following the Somogyi-Nelson method (37, 38). All assays were performed in triplicates.

2.7. Statistics and reproducibility

Whenever relevant biological or technical replicates were included in our study, and this information is provided in specific sections of the Methods chapter. All statistical tests used are indicated and the reference is provided. Correlation was calculated using Pearson correlation coefficient.

3. Results and discussion

3.1. Structural adaptation of termite gut microbiome to miscanthus diet

To examine enzymatic degradation of miscanthus by the higher termite gut system, two laboratory-maintained colonies (nests LM1 and LM3) of *Cortaritermes* intermedius were fed exclusively with dried miscanthus straw (Supplementary Fig. 1 and 2). This *Nasutitermitinae* genus is known to feed on grass tussocks in its natural habitat (39). Alteration of the termite gut microbiome (here relative to bacterial

communities in termite midgut and hindgut) was monitored at monthly basis during nine months, by high-throughput sequencing the V6-V8 regions of 16 S rRNA gene (Fig. 1a). Quality trimmed reads were assembled into 678 operational taxonomic units (OTUs) assigned to 18 bacterial phyla. *Spirochaetae* and *Fibrobacteres* were the most dominant, as previously shown for plant fibre-feeding higher termites (e.g. (11); Supplementary Data 1). By assessing bacterial community structures in control samples (colonies feeding on grass tussocks in situ) and miscanthus-fed microbiomes, we could observe radical changes. Species richness and diversity were significantly higher (HOMOVA $p<0.001$) before miscanthus diet was initiated, possibly reflecting an adaptive selection for the most efficient bacterial degraders facing lower complexity of carbon sources in comparison to original diet (Fig. 1b). Further application of linear discriminant analysis (LDA) effect size (LEfSe; (19)) to two termite colonies, demonstrated that nearly 140 bacterial OTUs were significantly enriched in control microbiome, while roughly 13 were enriched in miscanthus-fed microbiome (Supplementary Data 2). Out of the latter, six OTUs assigned to *Fibrobacteres* (mainly representing a genus exclusively containing termite *Fibrobaceteres* sequences) and *Spirochaetae* (associated with the termite *Treponema* cluster) were particularly abundant, and on average they accounted for 55.39 % ± 3.8 of the miscanthus-fed microbiome, in comparison to the 29.87 % ±1.8 average abundance in control microbiome. By analysing bacteria naturally associated with miscanthus diet, we estimated the effect of immigration on the termite gut community as negligible (Supplementary Figure 3). All together, these results demonstrated that diet change drives the development of bacterial consortium in a unique manner, yet food-associated bacteria cannot compete with highly specialised termite gut microbiota for a niche.

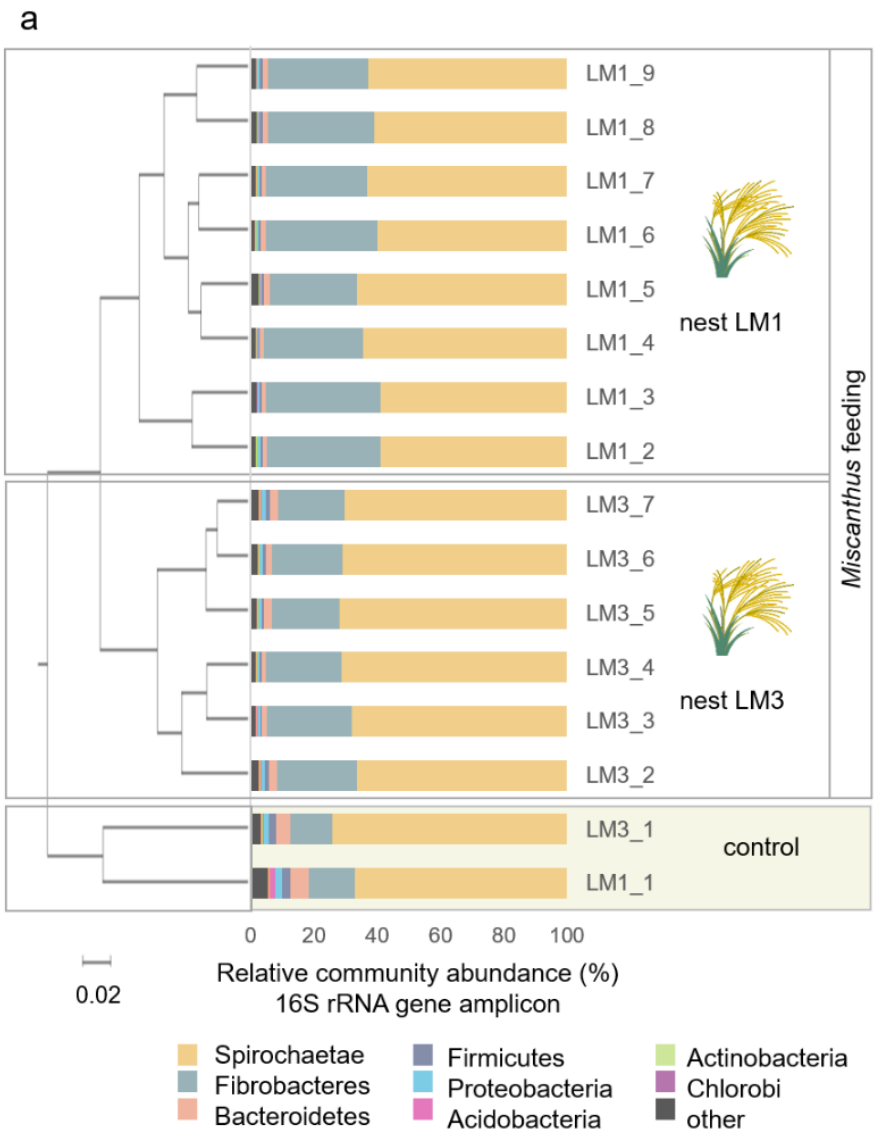


Figure 1a: Structural composition of *Cortaritermes* spp. gut microbiome under miscanthus diet. **(a)** - Clustering of samples based on the calculated Bray-Curtis index and phylum level taxonomic assignment of sequencing reads from the 16S rRNA gene amplicon study.

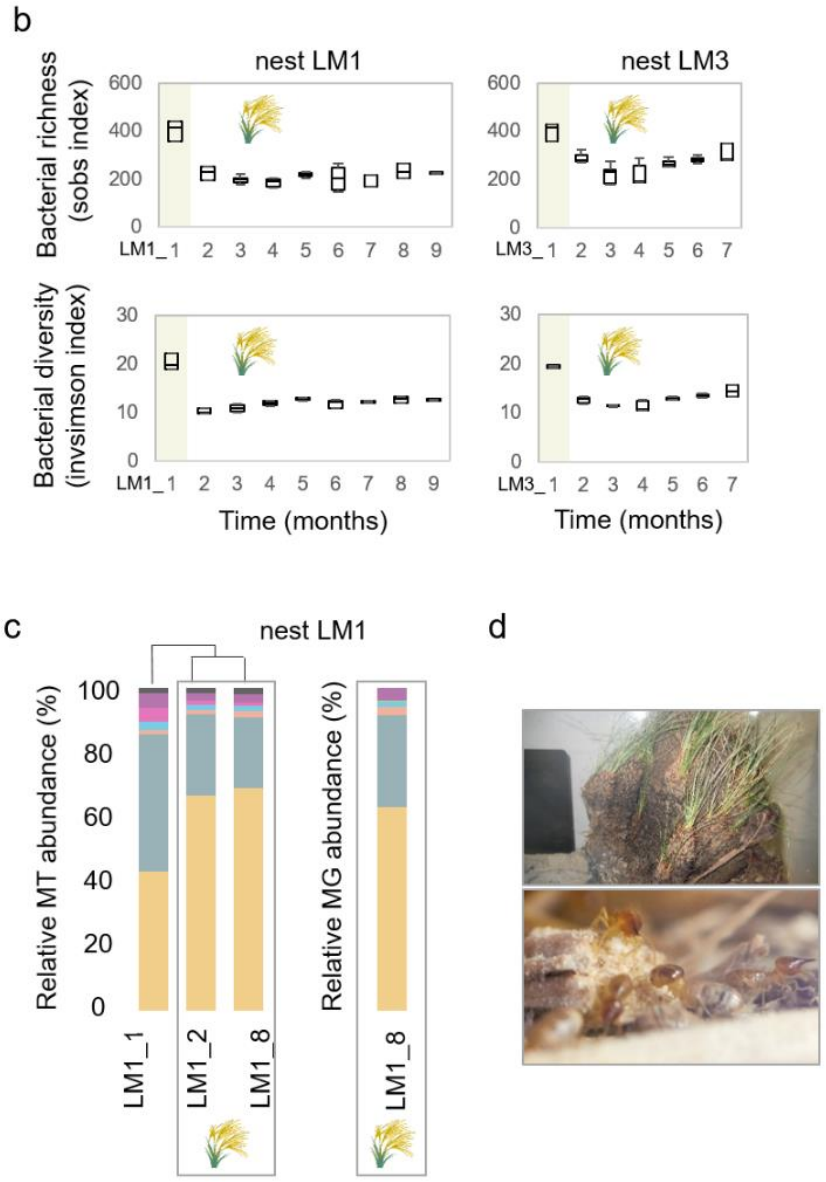


Figure 1b-c-d: Structural composition of *Cortaritermes* spp. gut microbiome under miscanthus diet. **(b)** - Bacterial richness and diversity indices before (highlighted in yellow on sub-figures a and b) and under miscanthus. Boxes represent the interquartile range and error bars show the 95% confidence intervals (n=3). **(c)** - Relative metatranscriptomic (MT) and metagenomic (MG) reads abundance assigned at the phylum level. Taxonomic gene and gene transcript assignments were inferred from the metagenomic contigs binning and phylum-level bin classification. **(d)** - *Cortaritermes* sp. colony (top) and termite workers under the protection of soldiers while feeding on miscanthus fibres in laboratory conditions (bottom).

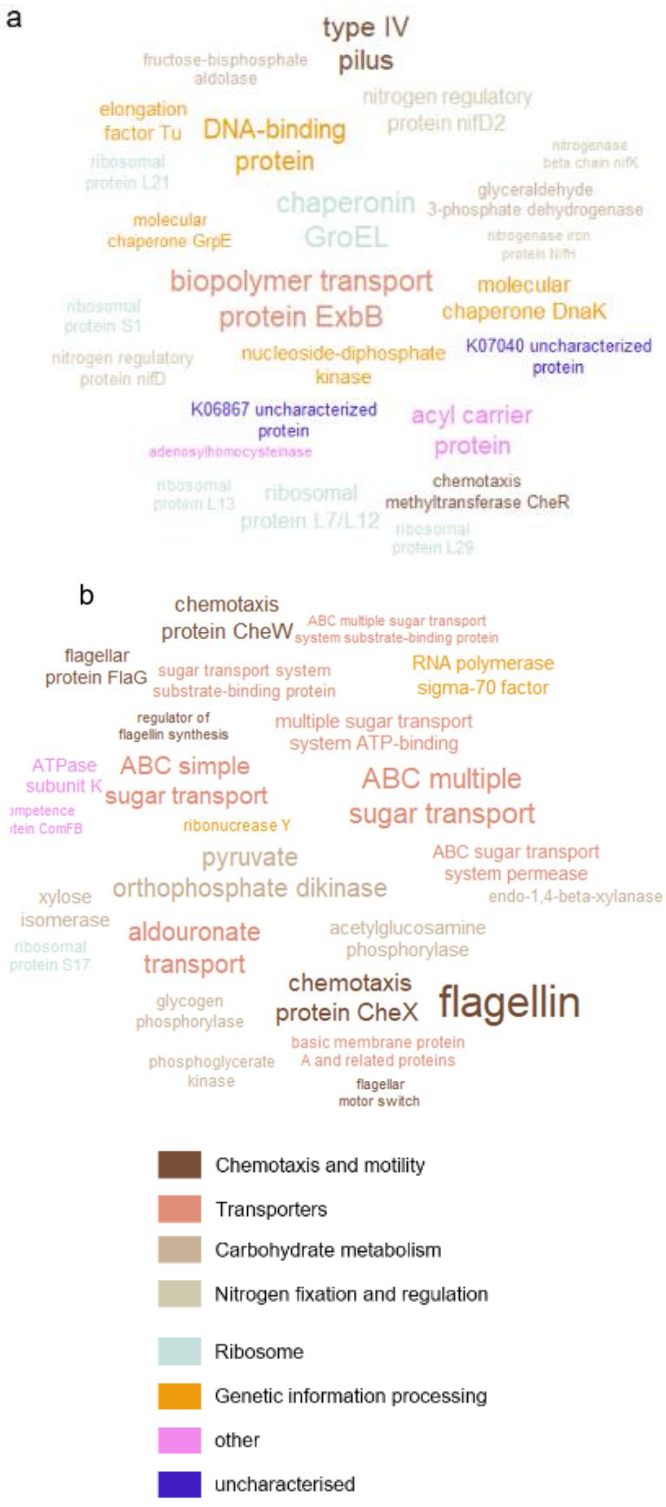
3.2. Comparison of *de novo* metatranscriptomic and metagenomic

The *de novo* MT was applied to nest LM1 and co-assembling reads from three samples (LM1_1 «control sample», LM1_2 and LM1_8, both representing « miscanthus-fed microbiomes») yielded 603,579 open reading frames (ORFs), mainly representing partially reconstructed gene transcripts. The *de novo* MG reconstruction was also applied to colony LM1 at sampling point LM1_8 and it yielded 211,724 ORFs annotated on 64,347 contigs for the total assembly size of 177,5 million base pairs (Mbp). For both datasets, initial public database-dependent taxonomic classification of genes and gene transcripts pointed to the abundance of *Firmicutes* (Supplementary Fig. 4), what contrasted the results of community structure analysis. Subsequent binning of MG contigs and phylum-level annotation of the resulting bacterial bins allowed assigning correct taxonomic origin, confirming the metagenomic abundance of *Spirochaetae* and *Fibrobacteres* (Fig. 1c; Supplementary Fig. 5). Following mapping of RNA reads to the MG assembly (referred to as "RNA-seq" analysis), we could confirm transcriptional dominance of these two bacterial phyla as well. Incomplete public databases and extensive horizontal gene transfer were previously proposed as the origin of this misclassification (9).

Based on the classification of genes and transcripts to broad functional categories such as KEGG ontology profiles (KOs), congruency between the *de novo* MG and MT reconstructions was high (Supplementary Fig. 6). However, out of the *de novo* MT reconstructed gene transcripts of prokaryotic origin only 37.8 % showed significant similarity to the *de novo* MG genes at the protein level (blastp e-value≤10-5), sharing on average 76.04 % of amino acid identity (Supplementary Fig. 4). Coherent to our study, differences in functional gene profiles between MG and MT reconstructions have been previously underlined (40). Even in the context of the termite gut, some authors highlighted the value of the *de novo* MT assembly in retrieving highly expressed genes (9, 41).

3.3. Genomic potential and transcriptional adaptation of gut bacteria

Aggregation of 68.9 ±1.8 % of the *de novo* reconstructed gene transcripts into clusters of orthologous genes (COGs) pointed at functional microbiome stability at the different stages of feeding campaign (Supplementary Data 3). Consistently with previous reports (11, 41, 42), cell motility and chemotaxis together with carbohydrate transport and metabolism were the two most highly expressed gene categories. Reconstruction of (nearly) complete metabolic modules was quite similar between *Fibrobacteres* and *Spirochaetae*. However, further comparative analysis using LEfSe (Fig. 2; Supplementary Data 4 and 5), identified several biologically informative features differentiating these two bacterial phyla. Both were capable of nitrogen fixation and glycogen synthesis, but the two pathways were enriched in *Fibrobacteres*. Expression of Amt ammonium transporters was highly up-regulated, and together with increased abundance of gene transcripts involved in urea transport and metabolism (restricted to *Spirochaetae*), it indicated nitrogen deficiency of a miscanthus fed termite colony. Both *Spirochaetae* and *Fibrobacteres* could also synthetize ten essential amino acids that animals cannot synthetize *de novo*. Even though, nitrogen provisioning by bacterial symbionts is not employed by all herbivorous insects, this strategy was proposed as a mechanism contributing to the success of termites (41) and herbivorous ants (43) in their marginal dietary niches. Importance of lignocellulose degradation under miscanthus diet was evidenced by increased abundance of transcripts broadly assigned to cellulose and xylan processing KOs (Fig. 2; Supplementary Data 4). Multiple sugar ABC transporters were up-regulated in the *Spirochaetae* metatranscriptome, while they were nearly absent from the MG and MT reconstructions of *Fibrobacteres* origin. This observation could suggest the governance of exogenous carbohydrates uptake and utilisation by *Spirochaetae*.


a
type IV pilus
fructose-bisphosphate aldolase
nitrogen regulatory protein nifD2
elongation factor Tu
DNA-binding protein
nitrogenase beta chain nifK
ribosomal protein L21
chaperonin GroEL
glyceraldehyde 3-phosphate dehydrogenase
molecular chaperone GrpE
nitrogenase iron protein NifH
biopolymer transport protein ExbB
molecular chaperone DnaK
ribosomal protein S1
nitrogen regulatory protein nifD
nucleoside-diphosphate kinase
K07040 uncharacterized protein
K06867 uncharacterized protein
acyl carrier protein
adenosylhomocysteinase
ribosomal protein L13
ribosomal protein L7/L12
chemotaxis methyltransferase CheR
ribosomal protein L29
b
chemotaxis protein CheW
ABC multiple sugar transport system substrate-binding protein
flagellar protein FlaG
sugar transport system substrate-binding protein
RNA polymerase sigma-70 factor
regulator of flagellin synthesis
multiple sugar transport system ATP-binding
ATPase subunit K
ABC simple sugar transport
ABC multiple sugar transport
ribonucrease Y
ABC sugar transport system permease
pyruvate orthophosphate dikinase
endo-1,4-beta-xylanase
xylose isomerase
aldouronate transport
acetylglucosamine phosphorylase
ribosomal protein S17
chemotaxis protein CheX
flagellin
glycogen phosphorylase
basic membrane protein A and related proteins
phosphoglycerate kinase
flagellar motor switch
Chemotaxis and motility
Transporters
Carbohydrate metabolism
Nitrogen fixation and regulation
Ribosome
Genetic information processing
other
uncharacterised

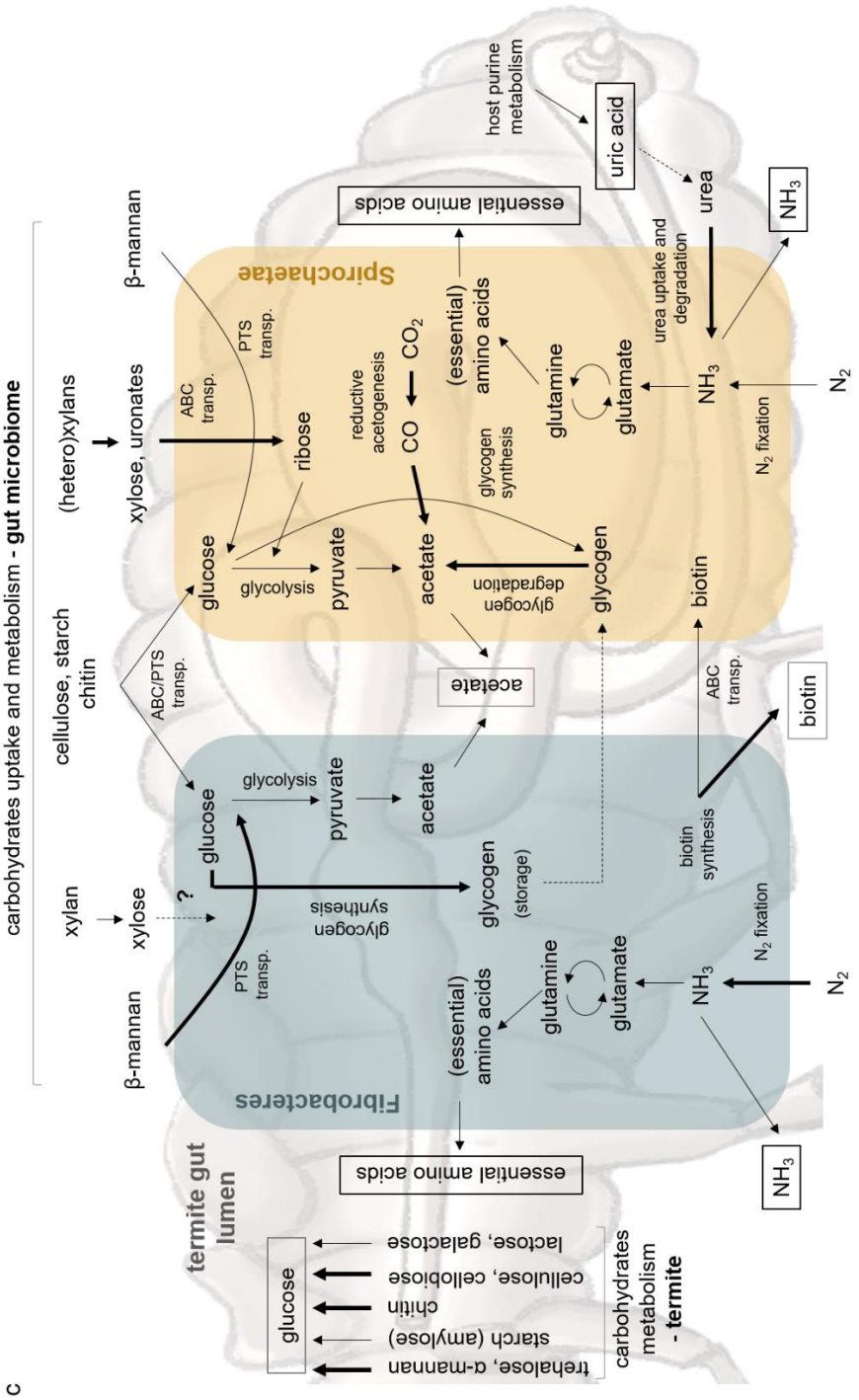

c
carbohydrates uptake and metabolism - gut microbiome
cellulose, starch
chitin
(hetero)xylans
xylose, uronates
β-mannan
xylan
xylose
β-mannan
ABC/PTS transp.
ABC transp.
PTS transp.
PTS transp.
?
glucose
glycolysis
pyruvate
acetate
glycogen synthesis
glycogen (storage)
glycogen degradation
ribose
reductive acetogenesis
CO
CO2
(essential) amino acids
glutamine
glutamate
NH3
N2
N2 fixation
biotin synthesis
ABC transp.
biotin
urea uptake and degradation
urea
host purine metabolism
uric acid
NH3
biotin
acetate
essential amino acids
Spirochaetae
Fibrobacteres
termite gut lumen
glucose
lactose, galactose
cellulose, cellobiose
chitin
starch (amylose)
trehalose, α-mannan
carbohydrates metabolism - termite

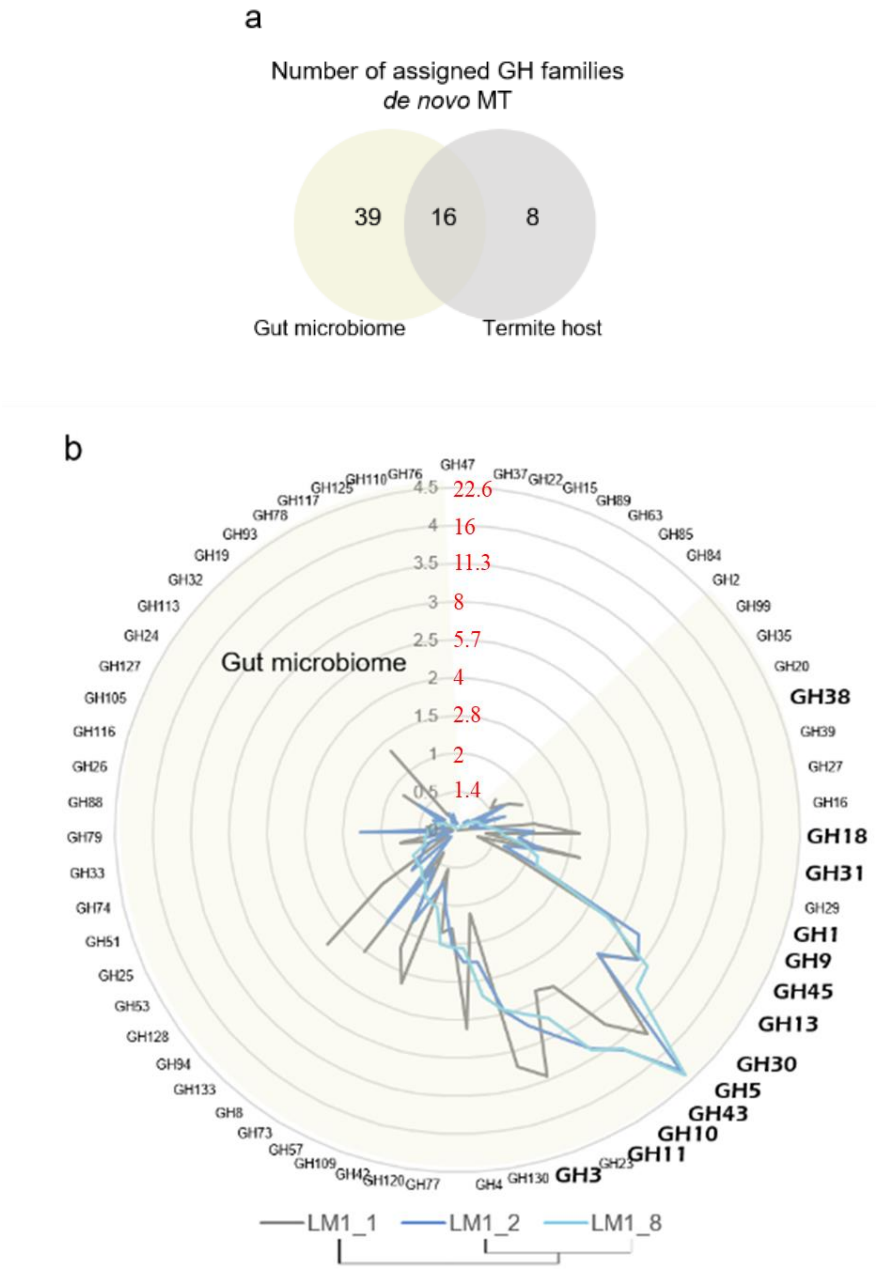
Figure 2: Functional characterisation of the termite gut system feeding on miscanthus. **(a, b)** - Tag clouds of enriched (LefSe LDA>2, p<0.05) KOs reconstructed from the *de novo* metatranscriptomics for the termite gut *Fibrobacteres* **(a)** and *Spirochaetae* **(b)** at LM1_8. Top 25 most abundant KOs are displayed. Size of the text reflects transcriptomics abundance of a specific KO. **(c) -** Simplified metabolic reconstruction, with a focus on carbohydrate metabolism, for the termite gut lignocellulolytic system. Hypothetical pathways are indicated with dashed lines. Metabolic pathways enriched in *Fibrobacteres* and *Spirochaetae* (metatranscriptomes) are indicated with bold lines. Metabolites putatively shared between gut bacteria and the host are indicated with square boxes.

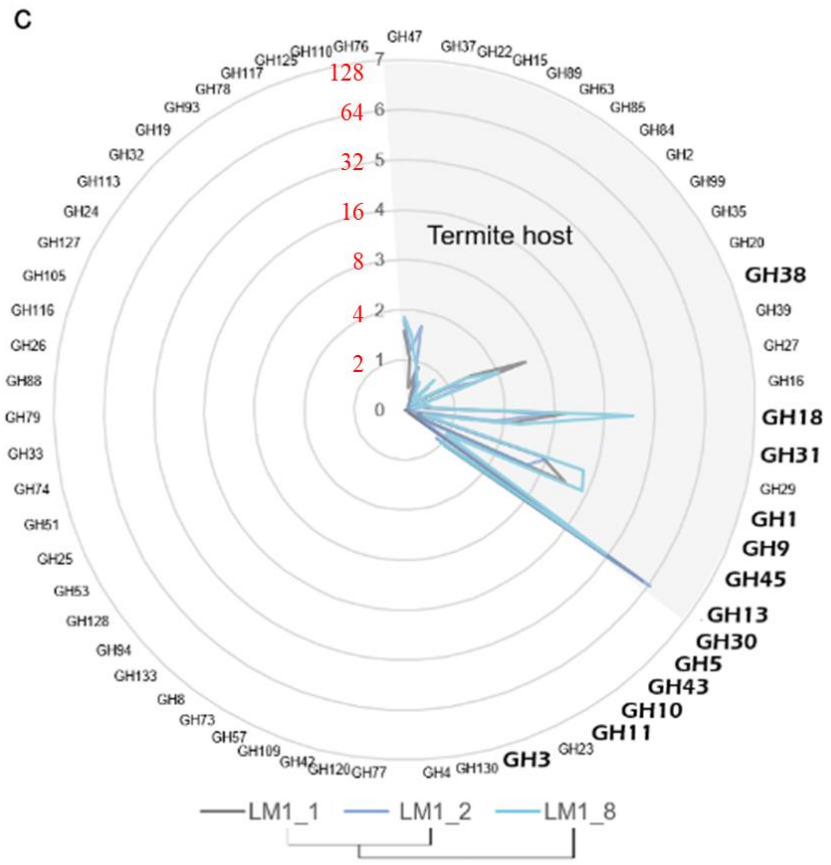
3.4. Diversity and abundance of termite gut bacterial CAZymes

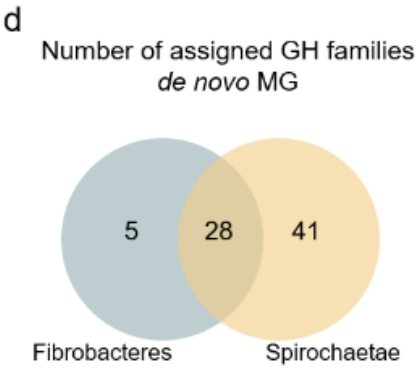
The *de novo* MT reconstruction yielded over 2000 of manually curated transcripts assigned as CAZymes-coding genes. Out of these, 38.4 % were further assigned to 55 GH families. The *de novo* MG reconstruction resulted in close to 7000 different CAZymes coding genes, 43.6 % of which were assigned to 86 GH families (Supplementary Fig. 6 and 7). Although there was a good correlation between the distributions of identified glycoside hydrolases to different GH families (Pearson r 0.83), roughly 150 genes coding for CAZymes identified in the *de novo* MT were also reconstructed from the assembled metagenome. Novelty of reconstructed CAZymes coding genes was evidenced through sequence comparison to NCBI refseq database, and a metagenomic dataset previously generated for *Nasutitermes* sp. (42)(Supplementary Fig. 4). In the latter case, average amino acids identity for the 943 query hits equalled 65.4 ± 19.9 % (blastp, e-value≤10-5), pointing to the diversity of bacterial CAZymes coding genes in guts of different termite species, even those phylogenetically closely related.

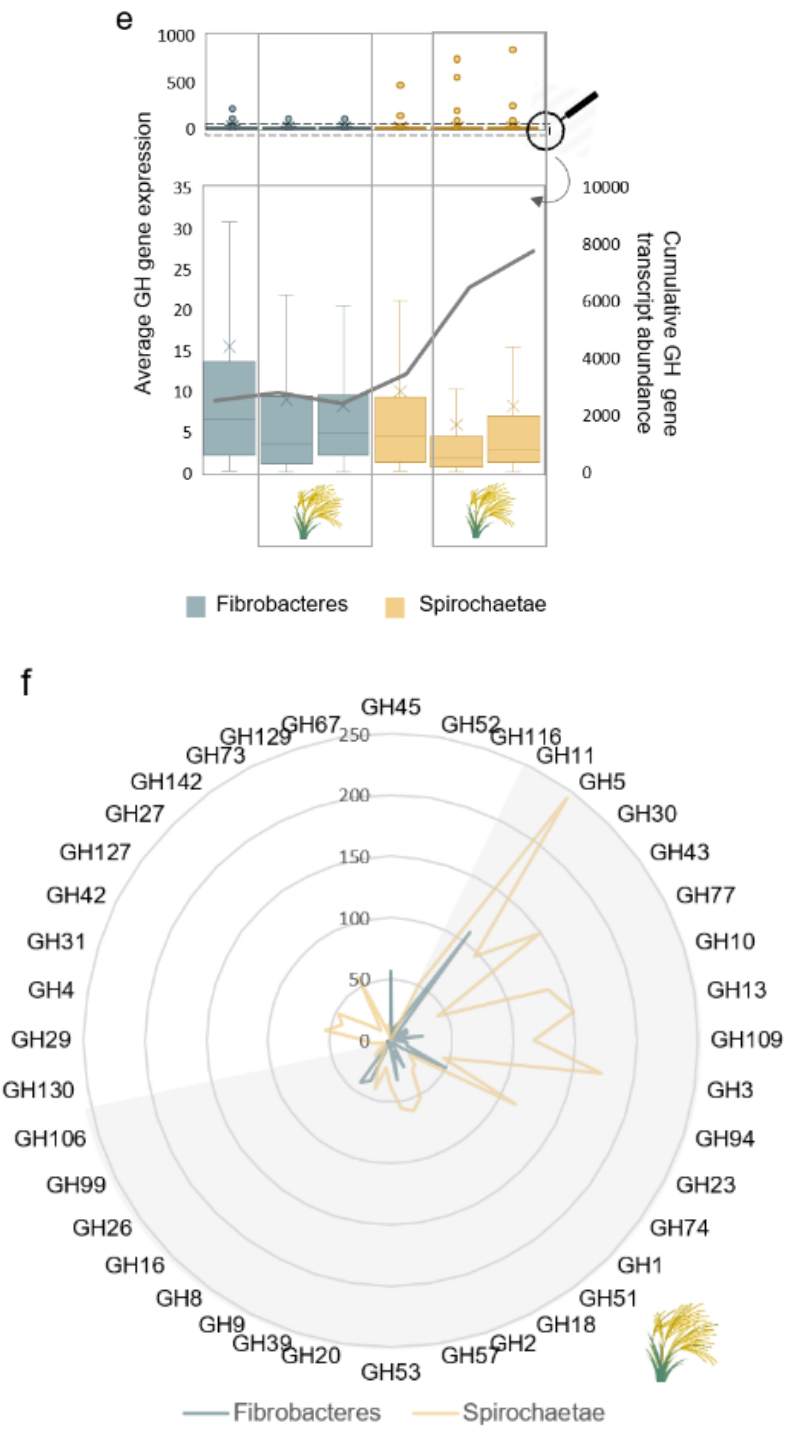
Differential expression of specific genes coding for CAZymes at different stages of the feeding experiment suggested quick acclimation to new (laboratory) conditions, also reflecting adaptation of gut bacteria to digest miscanthus (Fig. 3). There were roughly 29.7 % of common CAZymes coding genes transcripts between LM1_8 and control LM1_1 sample, while over 55 % were shared between LM1_8 and LM1_2 (both fed with miscanthus). Along the experiment, GH5 (mainly subfamilies GH5_4 and GH5_2) was the most highly expressed family. Still, its cumulative expression nearly doubled under miscanthus diet (Fig. 3b). Other abundant families included GH43, GH10 and GH11, all potentially involved in (hetero)xylan degradation. The

latter was previously shown as largely expressed by the termite gut fibre-associated *Spirochaetae* (9). Following manual curation, we removed three highly abundant but only partially reconstructed GH11 gene transcripts, what reduced initial over-dominance of this CAZy family by 3.3 ±0.9 fold (Supplementary Fig. 8). Similarly, only highly fragmented GH11 genes were recovered from the reconstructed metagenome. We hypothesise that closely related *Spirochaetae* strains contain highly similar GH11 genes, possibly shared by horizontal gene transfer, likewise it was shown for e.g. human gut *Bacteroidetes* (44). Their conserved nature may impede proper gene reconstruction from sequencing reads, similarly to other structurally conserved genes such as 16S rRNA or transposase, that typically do not reconstruct into larger genomic fragments (45) Indeed, the high similarity of such conserved genes within the metagenome, together with the generation of an important data volume and combined to the short read length, make the recovery of these conserved genes very difficult (46). GH45 was only present in *Fibrobacteres* and it was the second most expressed GH family under miscanthus diet. According to the CAZY database, nearly 95 % of proteins assigned to this family are of eukaryotic origin (Supplementary Fig. 10), and show endocellulase activity. Previously, GH45 CAZymes were characterised from lower termite symbiotic protists (47). Lytic polysaccharide monooxygenases (LPMOs) typically assigned to AA9 (fungal) and AA10 (predominantly bacterial enzymes) families (48) were neither present in the reconstructed metagenome nor metatranscriptome.


a
Number of assigned GH families
de novo MT
39
16
8
Gut microbiome
Termite host
b
Gut microbiome
4.5 22.6
4 16
3.5 11.3
3 8
2.5 5.7
2 4
1.5 2.8
1 2
0.5 1.4
GH47
GH37
GH22
GH15
GH89
GH63
GH85
GH84
GH2
GH99
GH35
GH20
GH38
GH39
GH27
GH16
GH18
GH31
GH29
GH1
GH9
GH45
GH13
GH30
GH5
GH43
GH10
GH11
GH23
GH3
GH130
GH4
GH77
GH120
GH42
GH109
GH57
GH73
GH8
GH133
GH94
GH128
GH53
GH25
GH51
GH74
GH33
GH79
GH88
GH26
GH116
GH105
GH127
GH24
GH113
GH32
GH19
GH93
GH78
GH117
GH125
GH110
GH76
LM1_1
LM1_2
LM1_8


c
Termite host
128
64
32
16
8
4
2
GH38
GH18
GH31
GH1
GH9
GH45
GH13
GH30
GH5
GH43
GH10
GH11
GH3
LM1_1
LM1_2
LM1_8


d
Number of assigned GH families
de novo MG
5
28
41
Fibrobacteres
Spirochaetae


e
1000
500
0
35
30
25
20
15
10
5
0
Average GH gene expression
10000
8000
6000
4000
2000
0
Cumulative GH gene transcript abundance
Fibrobacteres
Spirochaetae
f
GH45
GH52
GH116
GH11
GH5
GH30
GH43
GH77
GH10
GH13
GH109
GH3
GH94
GH23
GH74
GH1
GH51
GH18
GH2
GH57
GH53
GH20
GH39
GH9
GH8
GH16
GH26
GH99
GH106
GH130
GH29
GH4
GH31
GH42
GH127
GH27
GH142
GH73
GH129
GH67
250
200
150
100
50
0
Fibrobacteres
Spirochaetae

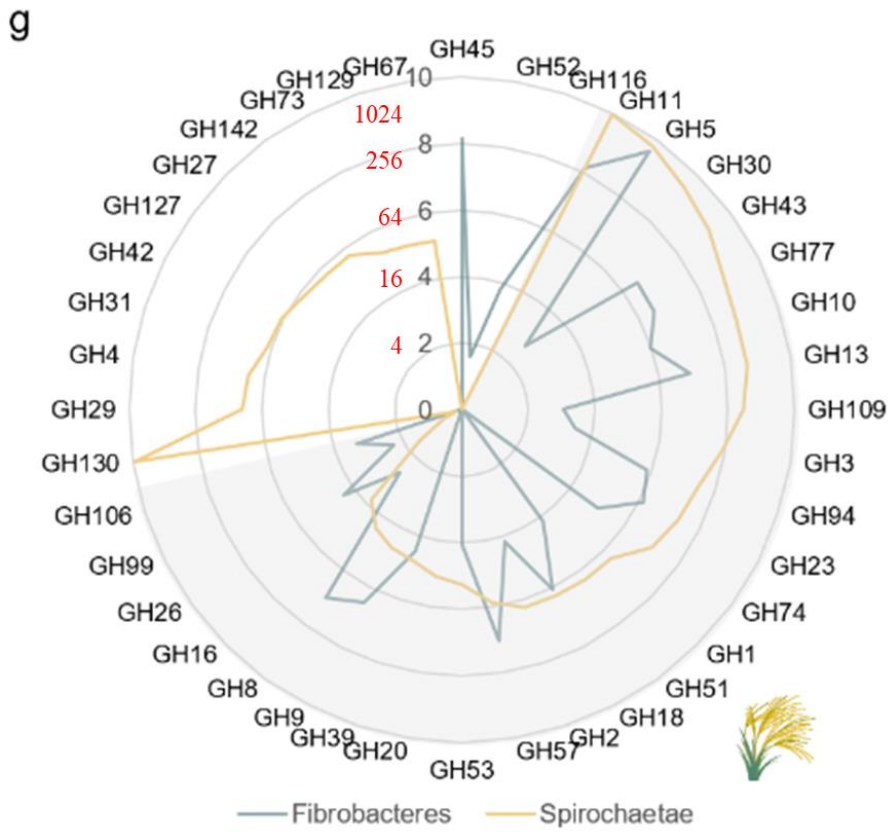


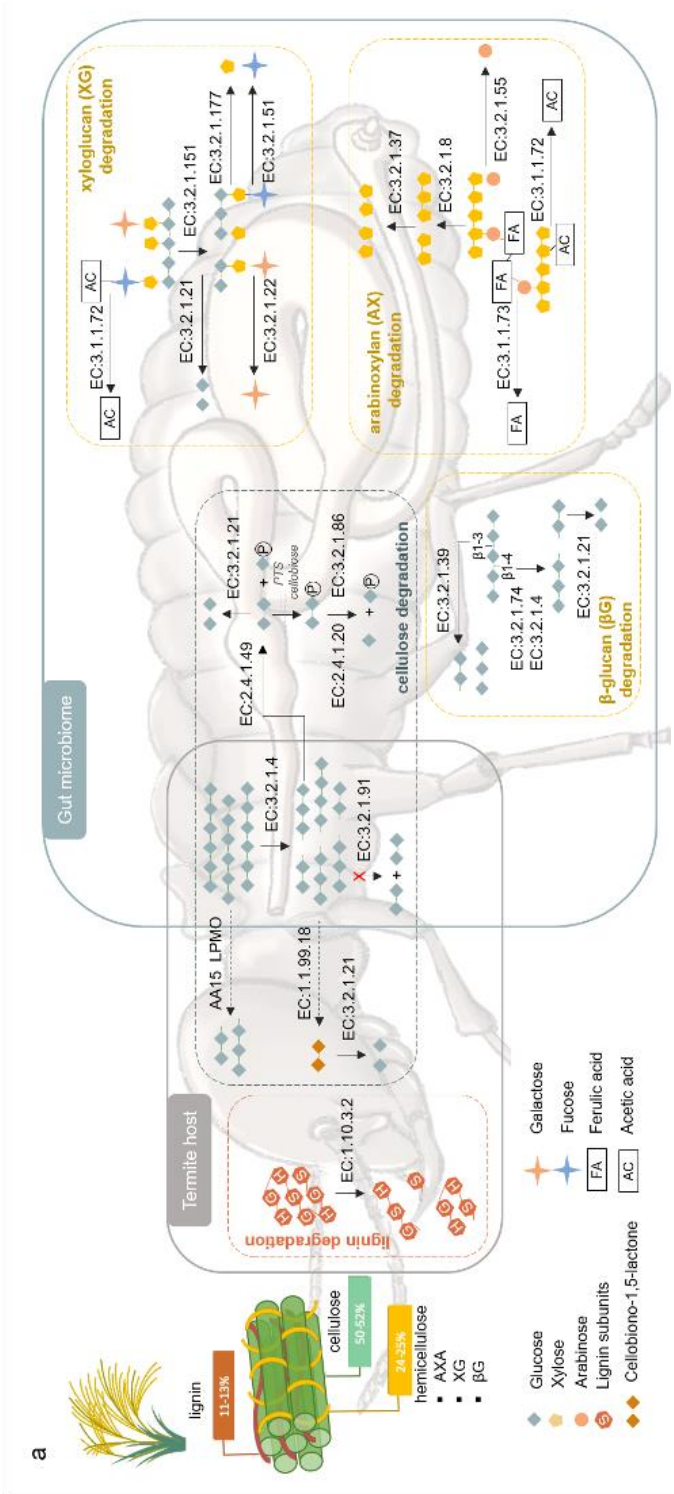
Figure 3: Carbohydrate active enzymes (CAZymes) reconstructed from metatranscriptomic (MT) and metagenomic (MG) reads for the termite gut system. **(a)** - Venn diagram showing the number of assigned glycoside hydrolase (GH) families for the termite gut microbiome and host gut epithelium. **(b,c)** - Comparison of the gene expression profiles (*de novo* MT, log2 transformed) for gene transcripts assigned to the different GH families and at the different stages of the miscanthus feeding experiment, for the gut microbiome **(b)** and the host gut epithelium **(c)** Corresponding transcripts per million (TPMs) values are presented in red. **(d)** - Venn diagram showing the number of assigned GH families for *Fibrobacteres* and *Spirochaetae*, based on the *de novo* MG reconstruction. **(e)** - Average CAZyme genes expression and cumulative expression of genes transcripts of the different GH families (in TPMs) at the different stages of the feeding experiment and analysed separately for *Fibrobacteres* and *Spirochaetae*. Lower panel is a zoom on the gene expression profiles with the outliers (highly expressed genes; in some cases representing only partially reconstructed genes) removed. Boxes represent the interquartile range and error bars show the 95% confidence intervals (n=number of transcripts annotated as glycoside hydrolases). **(f, g)** - Number of genes **(f)** and cumulative abundance of the most abundant GH families (**g**, RNA-seq log2 transformed) at the time point LM1_8 (the end of the miscanthus feeding experiment), and visualised separately for *Fibrobacteres* and *Spirochaetae*. Transcripts abundance **(g)** is calculated based on the RNA mappings (RNA-seq) to the MG contigs. Shaded parts correspond to shared GH families. Corresponding TPMs values are presented in red.

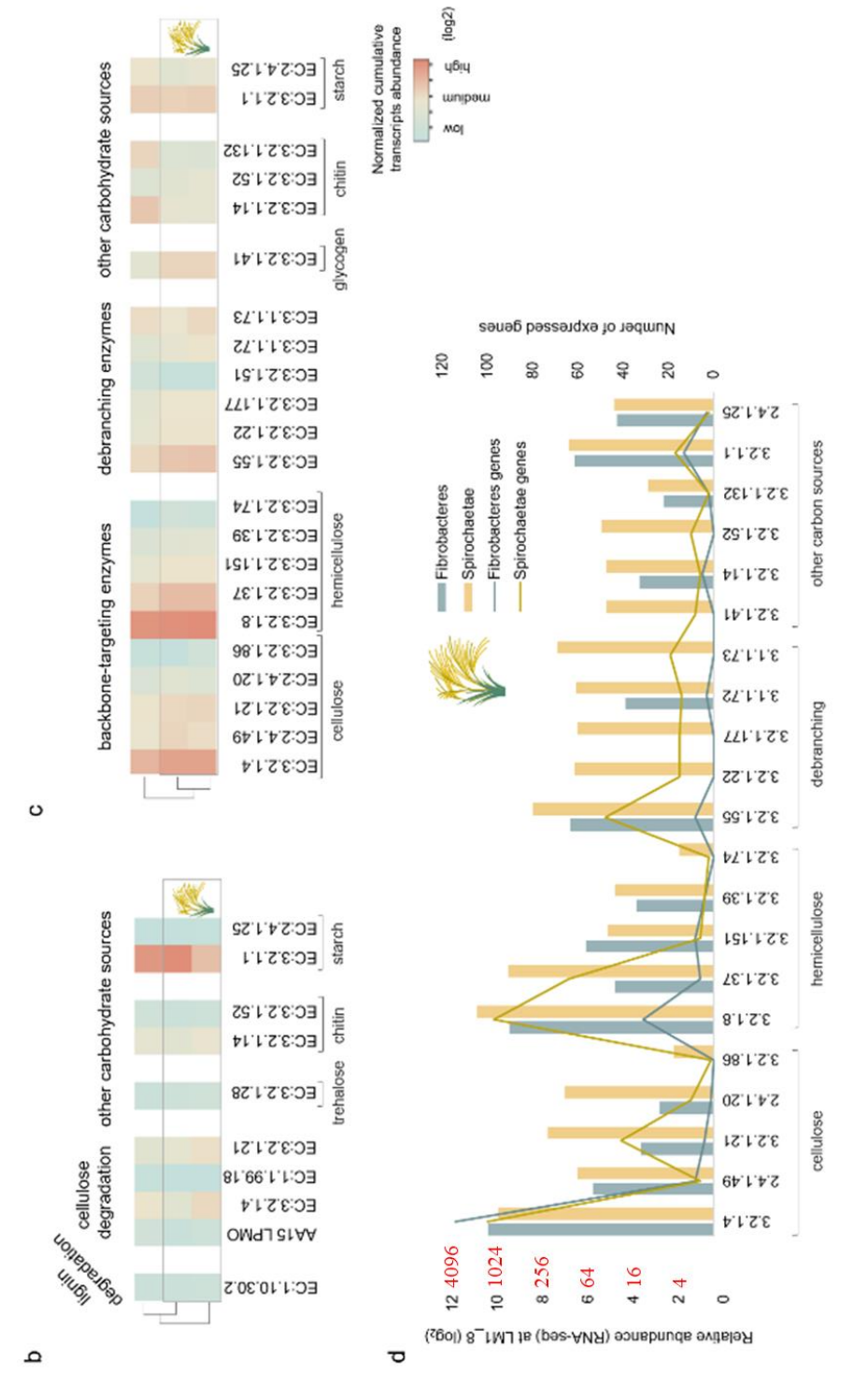
3.5. Expression and activities of GHs from *Fibrobacteres* and *Spirochaetae*

Based on the RNA-seq analysis, *Fibrobacteres* and *Spirochaetae* expressed respectively 47.9 ±15 % and 45.6 ±19 % of their CAZymes coding genes content when the termite was fed with miscanthus. Total diversity of *Spirochaetae* CAZymes coding genes was 2.3 ±0.5 fold higher than for *Fibrobacteres* (Fig. 3d-g). Their cumulative transcriptional abundance was also higher for *Spirochaetae*, however calculated average gene expression was slightly higher for *Fibrobacteres*. This observation was consistent across different GH families (Supplementary Fig. 11). As various GH families are characterised with broad functionalities, using peptide-based functional annotation (33) we further assigned *in silico* specific functions (EC numbers) to 60.3 ±2.2 % of gene transcripts classified as GHs. In many cases, these predictions were experimentally validated (Supplementary Data 6). We confirmed β-glucosidase, endocellulase, endoxylanase and arabinofuranosidase activities of several *Spirochaetae* CAZymes. We also characterised active endoxylanases and endomannanases from *Fibrobacteres*.

Abundance of transcripts associated with endocellulase (EC:3.2.1.4) and endoxylanase (EC:3.2.1.8) increased under miscanthus diet (for both *Fibrobacteres* and *Spirochaetae*), while those involved in chitin and starch (α-glucans) degradation decreased (Fig. 4, Supplementary Fig. 12). Endocellulase-assigned transcripts were nearly equally abundant between *Fibrobacteres* and *Spirochaetae*, while abundance and diversity of endoxylanases of *Spirochaetae* origin was much higher. Most of the assigned endocellulases were classified as GH5_4 enzymes (Supplementary Fig. 8.). Phylogenetic reconstruction comprising the previously characterised CAZymes from this family revealed the presence of multiple protein clusters separately grouping *Spirochaetae* and *Fibrobacteres* GHs (Fig. 5a). Concurrent inspection of reconstructed genomic fragments suggested the existence of different CAZymes coding genes loci containing GH5_4 genes (Supplementary Fig. 13). Interestingly, CAZymes previously characterised to possess single enzymatic activity (mostly endocellulase and to a lower extent endoxylanase) grouped in upper part of the tree. Lower part of the tree mainly contained multi-functional enzymes (single enzyme simultaneously acting on cellulose and xylan). Suggested enzymatic multi-

functionality was further confirmed for a selected GH5_4 CAZyme representing *Spirochaetae* cluster IX. Purified protein was shown to be an endocellulase acting on carboxymethylcellulose (CMC) and glucomannan (Fig. 5b). In addition, activity on xylan and arabinoxylan was confirmed. This gene was also one of the most highly expressed CAZymes coding genes under miscanthus diet, hypothesising the importance and interest for bacteria to express multi-functional enzymes. To our best knowledge, it represents first GH5_4 CAZyme of *Spirochaetae* origin ever characterised, and first multi-functional enzyme of higher termite gut prokaryotic origin.


a
lignin
11-13%
cellulose
50-52%
24-25%
hemicellulose
AXA
XG
βG
Termite host
Gut microbiome
lignin degradation
EC:1.10.3.2
AA15_LPMO
EC:1.1.99.18
EC:3.2.1.21
EC:3.2.1.4
EC:3.2.1.91
EC:2.4.1.49
EC:3.2.1.21
FTS
cellobiose
EC:2.4.1.20
EC:3.2.1.86
cellulose degradation
xyloglucan (XG) degradation
EC:3.1.1.72
AC
EC:3.2.1.21
EC:3.2.1.151
EC:3.2.1.177
EC:3.2.1.51
EC:3.2.1.22
arabinoxylan (AX) degradation
EC:3.2.1.37
EC:3.2.1.8
EC:3.2.1.55
EC:3.1.1.73
FA
EC:3.1.1.72
β-glucan (βG) degradation
EC:3.2.1.39
β1-3
EC:3.2.1.74
EC:3.2.1.4
β1-4
EC:3.2.1.21
Glucose
Xylose
Arabinose
Lignin subunits
Cellobiono-1,5-lactone
Galactose
Fucose
Ferulic acid
Acetic acid


b
lignin degradation
cellulose degradation
other carbohydrate sources
starch
chitin
trehalose
c
backbone-targeting enzymes
debranching enzymes
other carbohydrate sources
cellulose
hemicellulose
glycogen
chitin
starch
Normalized cumulative transcripts abundance
(log2)
low
medium
high
d
Relative abundance (RNA-seq) at LM1_8 (log2)
Number of expressed genes
Fibrobacteres
Spirochaetae
Fibrobacteres genes
Spirochaetae genes
cellulose
hemicellulose
debranching
other carbon sources

F**igure 4:** Characterisation of the termite gut lignocellulose degradation strategies. **(a)** - Simplified overview of enzymatic pathways involved in the degradation of main components of the miscanthus biomass, based on enzymes (gene transcripts assigned an EC number) revealed in our study. Dashed lines indicate hypothetical pathways. Lignin subunits correspond to: p-hydroxyphenyl (H), guaiacyl (G), and syringyl (S). CAZymes gene expression profiles (*de novo* metatranscriptomic) at the different stages of the miscanthus feeding experiment analysed for the termite gut epithelium **(b)** and the termite gut microbiome **(c)**. Gene expression analyses were done separately for the termite gut epithelium and the gut microbiome, therefore data on sub-figures b and c should only be compared within a single sub-figure. **(d)** - Relative CAZymes gene transcripts abundance (RNA-seq log2 transformed; the corresponding transcripts per million (TPMs) values are presented in red) and gene numbers assigned to different enzymatic categories and analysed separately for *Fibrobacteres* and *Spirochaetae* for LM1_8 sample.

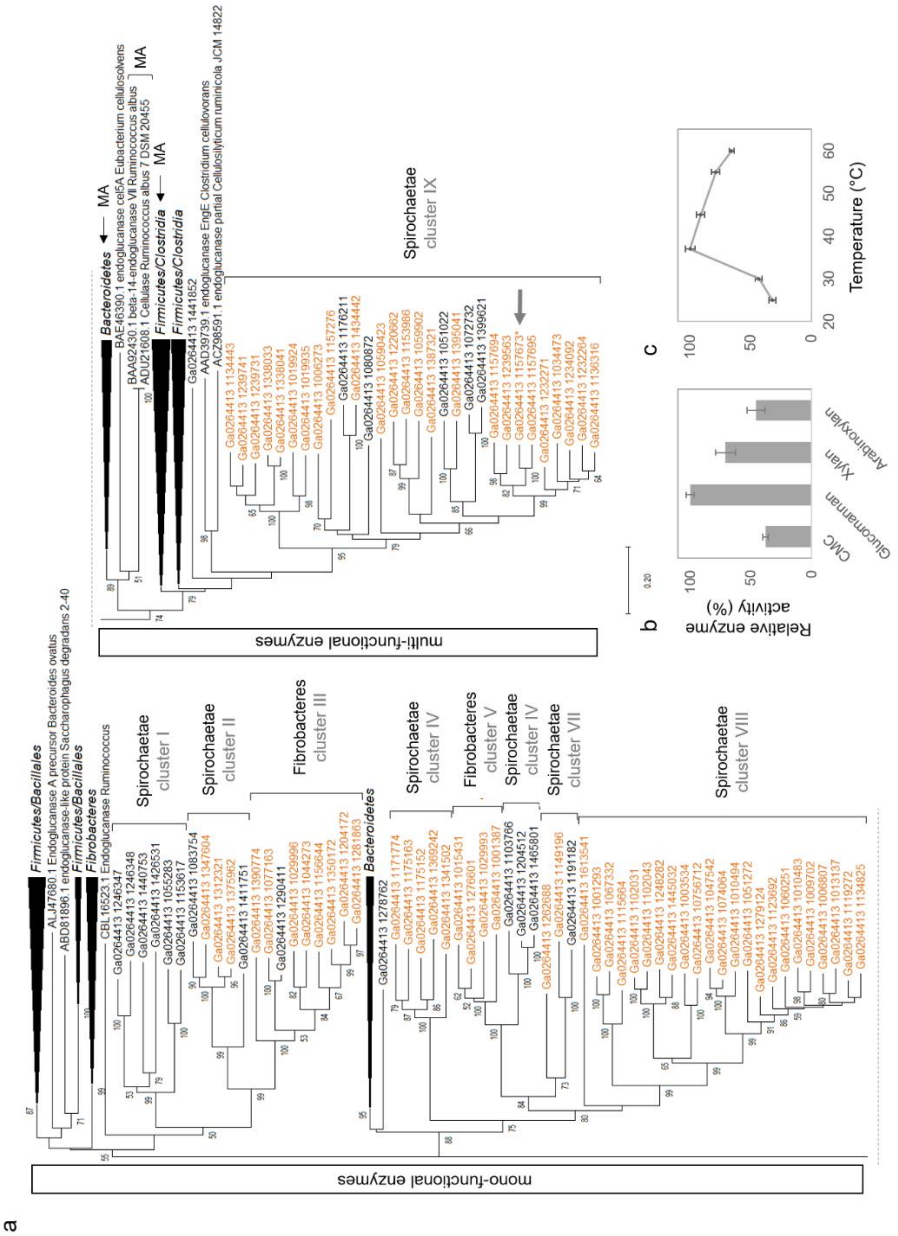

a
mono-functional enzymes
multi-functional enzymes
Firmicutes/Bacillales
Firmicutes/Bacillales
Fibrobacteres
Firmicutes/Bacillales
Bacteroidetes
Firmicutes/Clostridia
Firmicutes/Clostridia
Bacteroidetes
MA
MA
MA
Spirochaetae cluster I
Spirochaetae cluster II
Fibrobacteres cluster III
Spirochaetae cluster IV
Fibrobacteres cluster V
Spirochaetae cluster IV
Spirochaetae cluster VII
Spirochaetae cluster VIII
Spirochaetae cluster IX
b
0.20
c
Relative enzyme activity (%)
CMC
Glucomannan
Xylan
Arabinoxylan
Temperature (°C)

Figure 5: Characterisation of the GH5_4 family. **(a)** - Unrooted neighbor-joining tree containing the *de novo* reconstructed genes from metagenomic (MG) study (genes expressed under miscanthus diet are highlighted in orange on the tree). The tree was cut in two parts along the dashed line. All GH5_4 characterised proteins were retrieved from the CAZY database and included on the tree. Clusters indicated with an arrow and designated as "MA" contain known multi-functional enzymes. The percentage of replicate trees in which the associated sequences clustered together in the bootstrap test (500 replicates) are shown next to the branches. Final alignment involved 157 amino acid sequences. Protein from the *Spirochaetes* cluster IX indicated with a grey arrow was heterologously produced and characterised. **(b)** - Activity profiles for the heterologously produced and purified protein tested against CMC, glucomannan (galactomannan was negative, not displayed on the graph), xylan and arabinoxylan. **(c)** - Optimal temperature was assessed for glucomannan substrate. Error bars represent the standard deviation of a data set (n=3).

3.6. MAGs reconstruction and carbohydrate utilisation gene clusters

Reconstruction of draft bacterial genomes enriched in miscanthus-fed termite gut microbiome resulted in 20 MAGs with completeness above 50 % and contamination below 10 %, including eight *Fibrobacteres*-assigned MAGs, six *Spirochaetae* (four of *Treponema* origin) and three novel MAGs representing *Proteobacteria* phylum (Supplementary Data 7 and Supplementary Fig. 14). Average nucleotide identity (ANI) to MAGs from gut microbiomes of several higher termite species (49) equalled roughly 77.1 ±1.8 %, confirming the novelty of our MAGs. Frequency of CAZymes coding genes was higher in the *Treponema* genomes and they were also enriched in GHs. Over 36 % of annotated genes coding for CAZymes were aggregated into 1096 gene clusters containing more than one CAZy encoding gene. Similar gene clusters were recently discovered in gut microbiota of a wood-feeding termite *Globitermes brachycerastes* (50). Putative cellulose-utilisation gene clusters were the most highly expressed in *Fibrobacteres* MAGs, suggesting its major contribution to cellulose degradation (Fig. 3 and 4). Endo-xylanase-encoding genes were often co-localised with endocellulases (Supplementary Data 7), however xylosidases were scarce, questioning the ability of *Fibrobacteres* to utilise complex xylans. CAZymes coding genes clusters targeting among others alpha glucans, (arabino)xylans (AX), beta glucans, cellulose, chitin, galactans, mannans and xyloglucans were evidenced in reconstructed *Spirochaetae* genomes. Largely complete AX-targeting clusters were

assigned to MAG_17 and MAG_1, both representing *Treponema*. All of them were transcriptionally up-regulated under miscanthus diet (Supplementary Fig. 15), what is consistent with high AX content of miscanthus hemicellulose. Genes encoding for carbohydrate transporters (Supplementary Data 8) and in some cases even the two-component system response regulators and chemotaxis were adjacent to several CAZymes coding genes clusters. Organisationally similar to polysaccharide utilization loci (PULs) systems employed by *Bacteroidetes* (51), CAZymes coding genes clusters reconstructed in our study more resemble the concept of Gram-positive PULs recently proposed by Sheridan et al., (2016) in the context of human gut *Firmicutes* (52). Taken into account high sequence similarity of *Spirochaeteae* and *Firmicutes* CAZymes coding genes, it is possible that whole CAZymes coding genes clusters were acquired from *Firmicutes* in the course of evolution. Although not yet shown for *Spirochaetae*, *Bacteroidetes* PULs are often encoded within integrative and conjugative elements enabling their transfer among closely related species (44).

3.7. Host functional gene expression profiles under miscanthus diet

Following taxonomic annotation, 16,416 gene transcripts of eukaryotic origin were classified to *Arthropoda*. Based on the presence of 271 conserved orthologous reference eukaryotic genes (30), we estimated the completeness of our termite gut epithelial transcriptome (relative to midgut and hindgut) at 89.5 %, while the contamination with foreign mRNA was below 0.7 %. Number of assembled ORFs was in line with the two published termite genomes, *Macrotermes natalensis* (16,140 protein-coding genes (53)) and *Zootermopsis nevadensis* (15,459 protein coding-genes (54); Supplementary Data 9 and Supplementary Fig. 16). However, it was much lower than for another sequenced lower termite genome of *Cryptotermes secundus* (26,726 protein coding-genes; accession number PRJNA432597). Similarly to gut microbiome, gene transcripts related to carbohydrate transport and metabolism were abundant in the host transcriptome, showing importance of carbohydrates metabolism to the termite lifestyle (Supplementary Data 3). Over 10,000 transcripts were assigned a KO number and a summary of complete and partially reconstructed metabolic pathways is provided in Supplementary Data 4. Except for a previously partially reconstructed wood-feeding *Nasutitermes takasagoensis* transcriptome with 10,910

detected transcripts, our reconstruction represents the most complete transcriptome of a higher termite representing *Nasutitermitinae* subfamily (55).

We further identified 170 CAZY gene transcripts assigned to four main classes (GH, CE, AA and GT) and associated CBMs. Glycoside hydrolases were encoded by 66 genes, and their diversity patterns were similar to those identified in other termites with sequenced genomes (Supplementary Data 9). They were assigned to 19 different GH families, out of which five were not represented in the gut microbiome (Fig. 3a, c). The highest transcriptional abundance was attributed to GH13 (typically assigned as alpha glucanases; Supplementary Fig. 17), and it slightly decreased only towards the end of miscanthus campaign. Based on rather constant expression profiles of chitodextrinase, chitin utilization by the host did not change significantly upon miscanthus feeding (Supplementary Fig. 9). This indicates constant complementation of diet with nitrogen rich chitin, originating from either necrophagy and/or cannibalism, or fungi-colonised food stored in the nest, as previously proposed for other higher termite species (56). Transcriptional abundance of endocellulases increased at later stages of miscanthus feeding, suggesting a shift towards increased cellulose utilization by host. In addition, a gene transcript sharing 54 % of identity (at protein level) with a newly characterised cellulose- and chitin-targeting AA15 LPMO from *Thermobia domestica* (57) was also identified. This insect has a remarkable ability to digest crystalline cellulose without bacterial assistance. Further blast analysis revealed a presence of homologous genes in the other termite genomes, including *Z. nevadensis*, *M. natalensis* and *C. secundus*, suggesting that next to certain eukaryotes (e.g. crustaceans, molluscs, chelicerates, algae, and oomycetes) termites might be able to oxidatively cleave glycosidic bonds. Similar observation was also recently stated (58).

Interestingly, a few gene transcript of eukaryotic origin were not assigned to *Arthropoda*, suggesting the presence of active small eukaryotes in the higher termite gut. One gene transcript was assigned to AA3 family and classified as putative cellobiose oxidoreductase (EC 1.1.99.18; Fig. 4a, b). This type of oxidase is involved in oxidative cellulose and lignin degradation in wood-decaying fungi (59). In addition, two other transcripts assigned to GH45 family (putative endoglucanases) were present in the reconstructed transcriptome, and their expression slightly increased under miscanthus diet. Homology search (blastp) revealed their closest similarity to

Anaeromyces contortus, sp. nov. (*Neocallimastigomycota*), an anaerobic gut fungal species recently isolated from cow and goat faeces (60). In insects, GH45 CAZymes coding genes were previously detected in the genomes of Phytophaga beetles (61) and until now they were not reported from sequenced termite genomes.

3.8. Diet on miscanthus: who does what?

Miscanthus biomass is mainly composed of cellulose (41.4 ±2.9 %), hemicelluloses (25.8 ±5.2 %) including arabinoxylans , xyloglucans, β-glucans, and lignin (21.4 ±3.6 %) and other trace components (4). Based on the annotation of GH profiles and provided there is virtually no foreign nucleic acid contamination in the reconstructed termite transcriptome, termite on its own could digest amylose (starch), cellulose and/or cellobiose, lactose, galactose, chitin, mannan (α-mannan presumably contained in fungal cell wall) and mannose, trehalose, other glycans (e.g. N-acylsphingosine) and bacterial cell wall components (Fig. 4a, b). Until now, hemicellulose degradation seems to be conducted by gut bacteria, and no putative hemicellulolytic genes were recovered from the reconstructed termite transcriptome. For comparison, recent discovery of multifunctional GH9 cellulases in *Phasmatodea* suggested that some insects are capable to target xylan and xyloglucan in addition to cellulose (62). The same study, evidenced multifunctionality for two GH9 cellulases from *Mastotermes darwiniensis*, suggesting that some endogenous GH9 members in termites might also hydrolyse hemicellulose. Slightly increased transcriptional abundance of laccase-coding gene might indicate termite ability to target lignin. Beyond relatively high lignin content, recalcitrance of miscanthus biomass is mainly enhanced by other features, in particular acetylation and esterification with ferulic acid (FA). While acetyl esterase activity can be deduced from *Fibrobacteres* and *Spirochaetae* metatranscriptomes, the ability to break ferulic linkages seems limited to *Spirochaetae*. Putative feruoyl esterase from CE1 family are contained within AX-targeting CAZymes coding genes clusters in *Spirochaetae* MAGs, and were all up-regulated under miscanthus diet (Supplementary Fig. 15).

Based on the diversity and expression patterns of GHs, different sugar transporters (mainly ABC and to a lesser extent PTS; Supplementary Data 8) and specific sugar isomerases, *Spirochaetae* are able to utilise a wider range of miscanthus-derived sugars (including glucose, glucoronate, rhamnose, arabinose, mannose, xylose, ribose

and fucose) than *Fibrobacteres* (mainly glucose, mannose and possibly ribose). Both bacterial phyla can target the backbone of cellulose, xylans and mannans (the latter is not abundant in miscanthus biomass). Enrichment of *Fibrobacteres* metatranscriptome in endoglucanases (both targeting 1–3 β and 1–4 β glycosidic bonds as present in β-glucans and cellulose, respectively) shows its preference for carbohydrates with a glucose-unit backbone. *Fibrobacteres* also express endoxylanases, and we could confirm experimentally xylanase activity for one GH11 enzyme (Supplementary Data 6). However, hardly represented xylosidase-assigned gene transcripts and the absence of any xylose transporters and other known genes involved in xylose utilisation, would question the ability of *Fibrobacteres* to utilise xylans. Co-localisation of many endoxylanases together with potential endocellulases in the reconstructed *Fibrobacteres* MAGs further suggest that termite gut *Fibrobacteres* mainly remove xylan polymers from miscanthus fibres to better expose cellulose to the action of own endocellulases (Supplementary Fig. 13). By contrast, xylose isomerases and xylulose kinases were enriched in *Spirochaetae* metatranscriptome and both were highly expressed under miscanthus diet. All reconstructed gene transcripts were assigned to *Spirochaetae*, and together with enrichment of endoxylanase transcripts, it confirms the ability of these bacteria to degrade xylans, as recently proposed by Tokuda et al. (2018) (9).

Based on the *in silico* prediction of enzyme sub-cellular localizations, most of the endoxylanases from *Fibrobacteres* and *Spirochaetae* are exported outside the cell (Supplementary Data 10), suggesting initial degradation of xylan backbone in the extracellular space. Many *Spirochaetae* endocellulases are also putative extracellular enzymes or anchored to the outer membrane. In contrast, multiple *Fibrobacteres* endocellulases possibly lack signal peptide and are assumed to be localised in the cytoplasm. In general, as much as half of *Fibrobacteres* GHs are predicted to be localised in cytoplasm. At the same time, three times more GHs are exported outside the cell by *Spirochaetae*. This could indicate a rather selfish carbohydrates degradation strategy employed by termite *Fibrobacteres*, where cellulose fibres primarily detached by extracellular endocellulases are transported inside the cell for further breakdown. Selfish carbohydrate capture and degradation was previously proposed for *Bacteroidetes* in the rumen (63) and anaerobic digestion reactors (36). By contrast, recent work done on a rumen isolate *Fibrobacter succinogenes* S85

indicates that enzymes involved in cellulose degradation are localised on the cell surface (64). By maximizing intracellular cellulose breakdown, termite gut *Fibrobacteres* would avoid being in competition with much more abundant *Spirochaetae*. Enrichment of exbB (encoding for a biopolymer transport protein ExbB; Fig. 2a) gene transcripts in *Fibrobacteres* metatranscriptome would suggest possible cellulose/cellodextrin uptake through a mechanism similar to an experimentally demonstrated TonB-dependent transport of maltodextrins across outer membrane of *Caulobacter crescentus* (65), also discussed for *F. succinogenes* (64).

4. Conclusion and perspectives

Retrieving lignocellulose-active enzymes from naturally evolved biomass-degrading systems, with the use of continuously improving high throughput sequencing technologies, presents a promising strategy to identify new enzymes with potentially enhanced activities. Limited metatranscriptomic reports highlight high representation and overexpression of CAZymes in termite digestomes (e.g. (9, 41)). In higher termite guts, many lignocellulolytic steps are assisted by gut microorganisms (e.g. cellulose deconstruction), while some are exclusively attributed to hindgut bacteria (e.g. hemicellulose degradation). Cellulose degradation capacities of different lignocellulose degrading environments, including the termite gut system (66), have extensively been studied in the past. Decomposition of hemicellulose and general mobilisation of different lignocellulose components (breaking bonds between diverse plant polymers) have received comparably less scientific focus. Importantly, there is an increasing industrial interest in xylan-processing enzymes (67), regarding their application in biomass (wood) processing, pulp bio-bleaching, animal nutrition, food additives, etc. (68).

According to the recently established glycome profile of miscanthus, next to glucose, xylose and arabinose are the two main cell wall monosaccharides, both originating from arabinoxylan fibres which are ester-cross-linked by ferulic acid (69). Consequently, CAZymes coding genes specifically targeting (feruloylated) arabinoxylan components were highly up-regulated under miscanthus diet, making them potentially suitable candidates for industrial xylan-targeting applications. Many of these genes were found combined in clusters with a set of complementary hydrolytic activities to degrade e.g. AXs. Nature-optimised synergy between enzymes

of the same CAZyme cluster could further provide the basis to better define industrially relevant enzymatic cocktails. Specific lignocellulose fractions could be selectively targeted to deliver desired products, with potential effects being fully expectable and controllable (36). It would allow the fine-tuned deconstruction of lignocellulose for a variety of applications, e.g. oligoarabinoxylans for food industry (prebiotics), lignin fibres for biomaterials, glucomannans as food additives, etc. Feruloyl esterases, by removing cross-links between polysaccharides and lignin, help separating lignin from the rest of biomass, offering an alternative and/or complementation to currently applied industrial treatments (70). In addition, ferulic acid and other hydroxycinnamic acids can have many applications in food and cosmetic industries due to their antioxidant properties (71), thus further extending the application range of miscanthus biomass.

Approach-wise, experimental design undertaken in this study represents the enrichment strategy where a nature-derived microbial inoculum is grown in liquid batch cultures. Here, a natural system of the termite gut was shown to progressively adapt, yielding a consortium of microbes specialised in degradation of miscanthus biomass. Integrative omics combined with protein characterisation provides a framework for better understanding the complex lignocellulose degradation by the higher termite gut system and paves a road towards its future bioprospection.

Acknowledgements

This research has been funded through FNR 2014 CORE project OPTILYS (Exploring the higher termite lignocellulolytic system to optimize the conversion of biomass into energy and useful platform molecules/C14/SR/8286517), and co-funded through the grant PDR T.0065.15 from the Belgian F.R.S.-FNRS. We are grateful to Philippe Cerdan, Régis Vigouroux and the staff of the Laboratoire Environnement HYDRECO of Petit Saut (EDF-CNEH) for logistic support during the field work.

References

1. Y. M. Bar-On, R. Phillips, R. Milo, The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).
2. I. Lewandowski, J. C. Clifton-Brown, J. M. O. Scurlock, W. Huisman, Miscanthus: European experience with a novel energy crop. *Biomass and Bioenergy* **19**, 209–227 (2000).
3. H. Luo, *et al.*, Total Utilization of Miscanthus Biomass, Lignin and Carbohydrates, Using Earth Abundant Nickel Catalyst. *ACS Sustain. Chem. Eng.* **4**, 2316–2322 (2016).
4. T. van der Weijde, O. Dolstra, R. G. F. Visser, L. M. Trindade, Stability of cell wall composition and saccharification efficiency in Miscanthus across diverse environments. *Front. Plant Sci.* **7**, 1–14 (2017).
5. B. I. Cantarel, *et al.*, The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233–238 (2009).
6. A. Brune, Symbiotic digestion of lignocellulose in termite guts. *Nat. Rev. Microbiol.* **12**, 168–180 (2014).
7. S. E. Donovan, P. Eggleton, D. E. Bignell, Gut content analysis and a new feeding group classification of termites. *Ecol. Entomol.* **26**, 356–366 (2001).
8. A. Tartar, *et al.*, Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite Reticulitermes flavipes. *Biotechnol. Biofuels* **2**, 1–19 (2009).
9. G. Tokuda, *et al.*, Fiber-associated spirochetes are major agents of hemicellulose degradation in the hindgut of wood-feeding higher termites. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11996–E12004 (2018).
10. B. Godin, *et al.*, Chemical characteristics and biofuel potential of several vegetal biomasses grown under a wide range of environmental conditions. *Ind. Crops Prod.* **48**, 1–12 (2013).
11. M. Marynowska, *et al.*, Optimization of a metatranscriptomic approach to study the lignocellulolytic potential of the higher termite gut microbiome. *BMC Genomics* **18**, 681 (2017).
12. T. Miura, Y. Roisin, T. Matsumoto, Molecular phylogeny and biogeography of the nasute termite genus Nasutitermes (Isoptera: Termitidae) in the pacific tropics. *Mol. Phylogenet. Evol.* **17**, 1–10 (2000).
13. M. Calusinska, *et al.*, A year of monitoring 20 mesophilic full - scale bioreactors reveals the existence of stable but different core microbiomes in bio - waste and wastewater anaerobic digestion systems. *Biotechnol. Biofuels* **11** (2018).
14. A. Klindworth, *et al.*, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, 1–11 (2013).
15. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
16. P. Yilmaz, *et al.*, The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, 643–648 (2014).
17. P. D. Schloss, *et al.*, Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).

18. R Core Team., R: A language and environment for statistical computing. *Dendrochronologia* **0**, 1–16 (2020).
19. N. Segata, *et al.*, Metagenomic biomarker discovery and explanation. *genome Biol.* **12** (2011).
20. I. M. A. Chen, *et al.*, IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
21. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
22. H. H. Lin, Y. C. Liao, Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 12–19 (2016).
23. G. V. Uritskiy, J. DiRuggiero, J. Taylor, MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 1–13 (2018).
24. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
25. N. Segata, D. Börnigen, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4** (2013).
26. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).
27. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2018).
28. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
29. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
30. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
31. Y. Yin, *et al.*, DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, 445–451 (2012).
32. B. He, S. Jin, J. Cao, L. Mi, J. Wang, Metatranscriptomics of the Hu sheep rumen microbiome reveals novel cellulases. *Biotechnol. Biofuels* **12**, 1–15 (2019).
33. P. K. Busk, B. Pilgaard, M. J. Lezyk, A. S. Meyer, L. Lange, Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics* **18**, 1–9 (2017).
34. C. Savojardo, P. L. Martelli, P. Fariselli, G. Profiti, R. Casadio, BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* **46**, W459–W466 (2018).
35. A. S. Juncker, *et al.*, Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652–1662 (2003).
36. M. Bertucci, *et al.*, Carbohydrate Hydrolytic Potential and Redundancy of an

Anaerobic Digestion Microbiome Exposed to Acidosis , as Uncovered by Metagenomics. *Appl. Environ. Microbiol.*, 1–16 (2019).

37. M. Somogyi, A new reagent for the determination of sugars. *J. Biol. Chem.* **160**, 61–68 (1945).
38. N. Nelson, A photometric adaptation of the SOMOGYI method for the determination of glucose. *J. Biol. Chem.* **03**, 375–380 (1944).
39. C. Cuezzo, T. F. Carrijo, E. M. Cancello, Transfer of two species from NasutitermesDudley to CortaritermesMathews (Isoptera: Termitidae: Nasutitermitinae). *Austral Entomol.* **54**, 172–179 (2015).
40. A. Heintz-Buschart, P. Wilmes, Human gut microbiome: function matters. *Trends Microbiol.* **26**, 563–574 (2017).
41. S. He, *et al.*, Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. *PLoS One* **8** (2013).
42. K. Rossmassler, *et al.*, Metagenomic analysis of the microbiota in the highly compartmented hindguts of six wood- or soil-feeding higher termites. *Microbiome* **3**, 56 (2015).
43. Y. Hu, *et al.*, Herbivorous turtle ants obtain essential nutrients from a conserved nitrogen-recycling gut microbiome. *Nat. Commun.* **9** (2018).
44. M. J. Coyne, N. L. Zitomersky, A. M. McGuire, A. M. Earl, L. E. Comstock, Evidence of extensive DNA transfer between bacteroidales species within the human gut. *MBio* **5**, 1–12 (2014).
45. D. H. Parks, *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2** (2017).
46. C. Yuan, J. Lei, J. Cole, Y. Sun, Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* **31**, i35–i43 (2015).
47. J. P. L. Franco Cairo, *et al.*, Expanding the knowledge on lignocellulolytic and redox enzymes of worker and soldier castes from the lower termite Coptotermes gestroi. *Front. Microbiol.* **7** (2016).
48. S. M. Cragg, *et al.*, Lignocellulose degradation mechanisms across the Tree of Life. *Curr. Opin. Chem. Biol.* **29**, 108–119 (2015).
49. V. Hervé, *et al.*, Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. *PeerJ* **2020**, 1–27 (2020).
50. N. Liu, *et al.*, Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *ISME J.* (2018) https:/doi.org/10.1038/s41396-018-0255-1.
51. M. Zhang, *et al.*, Xylan utilization in human gut commensal bacteria is orchestrated by unique modular organization of polysaccharide-degrading enzymes. *Proc. Natl. Acad. Sci.* **111**, E3708–E3717 (2014).
52. P. O. Sheridan, *et al.*, Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic firmicutes. *Microb. Genomics* **2**, 1–16 (2016).
53. M. Poulsen, *et al.*, Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 14500–14505 (2014).
54. N. Terrapon, *et al.*, Molecular traces of alternative social organization in a termite genome. *Nat. Commun.* **5** (2014).

55. R. P. Kumara, S. Saitoh, H. Aoyama, N. Shinzato, G. Tokuda, Metabolic pathways in the mixed segment of the wood-feeding termite Nasutitermes takasagoensis (Blattodea (Isoptera): Termitidae). *Appl. Entomol. Zool.* **51**, 429–440 (2016).
56. L. Menezes, *et al.*, Food Storage by the Savanna Termite Cornitermes cumulans (Syntermitinae): a Strategy to Improve Hemicellulose Digestibility? *Microb. Ecol.* **76**, 492–505 (2018).
57. F. Sabbadin, *et al.*, An ancient family of lytic polysaccharide monooxygenases with roles in arthropod development and biomass digestion. *Nat. Commun.* **9** (2018).
58. G. Tokuda, *Advances in Insect Physiology Vol. 57*, J. Russell, Ed. (Academic Press, 2019).
59. R. Berlemont, Distribution and diversity of enzymes for polysaccharide degradation in fungi. *Sci. Rep.* **7**, 222 (2017).
60. R. A. Hanafy, B. Johnson, M. S. Elshahed, N. H. Youssef, Anaeromyces contortus, sp. Nov., a new anaerobic gut fungal species (neocallimastigomycota) isolated from the feces of cow and goat. *Mycologia* **110**, 502–512 (2018).
61. A. Busch, E. G. J. Danchin, Y. Pauchet, Functional diversification of horizontally acquired glycoside hydrolase family 45 (GH45) proteins in Phytophaga beetles. *BMC Evol. Biol.* **19**, 1–14 (2019).
62. M. Shelomi, B. Wipfler, X. Zhou, Y. Pauchet, Multifunctional cellulase enzymes are ancestral in Polyneoptera. *Insect Mol. Biol.* **29**, 124–135 (2020).
63. L. M. Solden, *et al.*, Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).
64. M. P. Raut, N. Couto, E. Karunakaran, C. A. Biggs, P. C. Wright, Deciphering the unique cellulose degradation mechanism of the ruminal bacterium Fibrobacter succinogenes S85. *Sci. Rep.* **9**, 1–15 (2019).
65. H. Neugebauer, *et al.*, ExbBD-Dependent Transport of Maltodextrins through the Novel MalA Protein across the Outer Membrane of Caulobacter crescentus. *J. Bacteriol.* **187**, 8300–8311 (2005).
66. A. Brune, "Microbial Symbioses in the Digestive Tract of Lower Termites" in *Beneficial Microorganisms in Multicellular Life Forms*, E. Rosenberg, U. Gophna, Eds. (Springer Berlin Heidelberg, 2011), pp. 3–25.
67. M. Sharma, A. Kumar, Xylanases: An Overview. *Br. Biotechnol. J.* **3**, 1–28 (2013).
68. A. Walia, S. Guleria, P. Mehta, A. Chauhan, J. Parkash, Microbial xylanases and their industrial application in pulp and paper biobleaching: a review. *3 Biotech* **7**, 1–12 (2017).
69. R. M. F. da Costa, *et al.*, A cell wall reference profile for Miscanthus bioenergy crops highlights compositional and structural variations associated with development and organ origin. *New Phytol.* **213**, 1710–1725 (2017).
70. A. E. Fazary, Y. H. Ju, Feruloyl esterases as biotechnological tools: Current and future perspectives. *Acta Biochim. Biophys. Sin. (Shanghai).* **39**, 811–828 (2007).
71. A. Dilokpimol, *et al.*, Diversity of fungal feruloyl esterases : updated phylogenetic classification , properties , and industrial applications. *Biotechnol. Biofuels* **9**, 1–18 (2016).

Chapter 5

General discussion, conclusions and perspectives

While lignocellulose is regarded as a renewable source of bioenergy and of various commodity chemicals that could support the development of the future biorefinery sector (1), it is often highly recalcitrant (resistant to enzymatic hydrolysis) and therefore its enzymatic deconstruction is challenging. Consequently, the use of lignocellulosic biomass for the biorefinery sector requires environmentally friendly, efficient and cost-effective pre-treatments. However, currently available pre-treatment processes (including physico-chemical pre-treatments) are energy demanding and/or can lead to the production of unwanted by-products (*e.g.* furfurals derived from acid treatment of hemicellulose), that inhibit or hinder subsequent processes in the biorefinery value chain (2, 3). An optimal pre-treatment (here defined as a strategy to deconstruct lignocellulose biomass into simpler/simple components) should minimise biomass recalcitrance in a more specific and cost-efficient way, at the same time avoiding the formation of unwanted by-products (*e.g.* furfurals). Moreover, selective biomass conversion, specifically targeting its different fractions, would leave the integrity of fermentable sugars/oligosaccharide products untouched (without significant chemical modifications), thus making them suitable for direct or indirect production of *e.g.* industrial commodities. This would allow for a complete valorisation of plant biomass, including hemicellulose as well as lignin, which is the only naturally occurring aromatic polymer, currently underutilised (4).

Although the biorefinery of biomass is a man-made concept that does not occur in nature, natural organisms/systems can effectively mediate the different steps in the course of the process. In nature, next to some fungi and bacteria, lignocellulose deconstruction can be performed by a restricted number of animals that in most cases rely on the synergistic actions of their syntrophic microorganisms. These animals can produce their own carbohydrate active enzymes, e.g. cellulases, mannanases, etc. (5), able to cleave glycosidic bonds. However, they mainly rely on specific fungi and bacteria to deconstruct or detach cellulose and hemicellulose from the lignin structure (*Chapter 1*). As a good example of such a natural lignocellulose degrading system, wood feeding higher termite gut system can convert up to 74-99% and 65-87% of cellulose and hemicellulose, respectively, from the ingested biomass (6). Another example is the microbe-driven but man-optimised processes of anaerobic digestion, which has been shown to efficiently deconstruct some of the recalcitrant

lignocellulosic biomasses as well (7). Still, pretreatment of the biomass substrates is often required in order to achieve an important yield of deconstruction in the bio-reactor.

Therefore, better understanding of microbe-driven lignocellulose deconstruction strategies, by studying natural systems such as the termite gut, and man-engineered systems such as the anaerobic digestion process, has the potential to reveal eco-friendly alternatives for hydrolytic deconstruction of lignocellulosic biomass. To that purpose, with the use of novel metagenomics and metatranscriptomics approaches, I investigated these two biomass deconstructing systems. First, relying on the output of the applied omics technologies, I characterized bacterial enzymatic strategies in terms of lignocellulose deconstruction, and I identified *in silico* the most interesting enzymes with lignocellulolytic activities (*Chapters 2, 3 and 4*). Second, I cloned respective genes and overproduced them in heterologous hosts (*Chapters 2 and 4*). Third, I purified produced enzymes and assessed their hydrolytic activities on different lignocellulose components, sometimes combining them in enzymatic cocktails. e.g. glucomannan-degrading polysaccharide utilization locus (*Chapter 2*). The next step would be to verify their applicability in a hydrolytic pre-treatment of complex biomasses, that was only very partially covered during my PhD, using a miscanthus biomass as an example (results not included into this document due to their very preliminary nature).

1. Combination of metagenomics and metatranscriptomics allows for the identification and characterization of key bacterial players involved in lignocellulose deconstruction

Combined metagenomic and metatranscriptomic approach targeting bacteria allows the identification of less abundant genes but highly expressed transcripts, and therefore important enzymes involved in lignocellulose deconstruction.

Using high throughput DNA sequencing technologies (metagenomics) in order to retrieve lignocellulose-active enzymes from naturally evolved biomass-degrading bacterial communities, presents an interesting strategy to identify new enzymes with

putative improved activities. In this context, Hess et al. (8) previously identified 27755 putative carbohydrate-active enzymes from the 268Gb of the cow rumen metagenomic data. As indicated by a rarefaction analysis, only a subset of genes present in the cow rumen microbiome was assembled, despite the considerable sequencing depth of the study. The average frequency of target genes in bacterial genomes is lower than two cellulolytic glycoside hydrolase per bacterial genome (9). Consequently, the risk to omit important cellulolytic activity from the metagenomic analysis is considerably higher than through an mRNA approach (metatranscriptomics), which potentially reflects the effort of the system to break down lignocellulose into useful components.

Previous metatranscriptomic reports highlighted the high representation and overexpression of cellulose and hemicelluloses deconstructing genes in the termite hindgut digestomes (10–12) and are in line with the discoveries presented in this PhD thesis. Similarly to our results, previous metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding termites (12) highlighted the carbohydrate transport and metabolism to be one of the most expressed functional gene categories, again indicating the potential of metatranscriptomics to discover new carbohydrate-active enzymes in such an environment (12). Interestingly, these authors highlighted the value of the *de novo* metatranscriptome assembly in retrieving highly expressed genes (e.g. the top ten highly expressed glucose hydrolases were identified solely from the *de novo* assembled metatranscriptome), that were otherwise present at low relative abundance in metagenomes. The importance of the *de novo* transcriptome assembly to reconstruct and functionally characterize abundant transcripts that were underrepresented in the corresponding metagenomic data set was also raised in our study (*Chapter 5*). We clearly highlighted that less than half of *de novo* reconstructed gene transcripts (metatranscriptomics sequencing) showed similarity to reconstructed genes (metagenomics) in the study of the *Cortaritermes* feeding on miscanthus biomass (*Chapter 5*). Moreover, roughly 150 out of 2000 *de novo* reconstructed carbohydrate active enzymes gene transcripts were also retrieved in the metagenome. While these differences could have partially arisen from the medium output of next generation sequencing applied in our study, they also reflect the functional plasticity of bacterial communities, which implies that bacterial players have the potential to

adapt to perturbations by modulating gene expression, what also leads to the high discrepancy between the two datasets (13).

Combined metagenomic and metatranscriptomic approach allows the identification of unexpected phylum with high lignocellulolytic potential.

The advantage of the complementarity of both approaches was further evidenced in the study focusing on the continuously stirred tank anaerobic reactors fed with sugar beet pulp (*Chapter 3*). Based on the diversity of reconstructed carbohydrate active enzymes from this metagenomics study, I evidenced the unexpected high lignocellulose deconstructing potential of the *Planctomycetes* phylum (*Chapter 3*, Figures 2 and 3), far exceeding these of *Bacteroidetes*, *Firmicutes* and *Chloroflexi*, which are commonly regarded as hydrolytic actors in this environment. However, in anaerobic digestion reactors, *Planctomycetes* do not take a full advantage of their hydrolytic capacity, as only a minor fraction of encoded carbohydrate active enzymes was actually expressed, as shown by the complementary metatranscriptomics study (*Chapter 3*, Figure 4). In turn, in the case of *Bacteroidetes* and to a lower extent for *Firmicutes* and *Chloroflexi*, most of the re-constructed carbohydrate active enzymes were expressed, suggesting their active involvement in biomass deconstruction. Still, taken the abundance of carbohydrate active enzymes in sequenced *Planctomycetes* genomes, I hypothesised that these bacteria might be important but currently overlooked lignocellulose degraders in biomass-rich environments. Based on my results that evidenced the high diversity of encoded glycoside hydrolase families in a single *Planctomycetes* genome, I postulated that particular bacterial isolates could target a whole range of lignocellulose components, thus opening the door to their application in the developing bio-refinery sector. Interestingly, this outcome of my PhD thesis will be further exploited by the former research team in a context of a recently acquired research project.

Combined metagenomic and metatranscriptomic approach allows the characterization of new bacteria together with new clusters of genes targeting lignocellulose deconstruction.

While metagenomics allows describing the metabolic potential of a bacterial community or a specific organism forming part of such community,

metatranscriptomics characterises bacteria actively participating to the observed community trait (13), as evidenced above taking the example of *Planctomycetes*. The application of the metagenomics sequencing combined with metagenome assembled genomes reconstruction, including the work done in the frame of this PhD thesis (*Chapters 2 and 3*), recently led to the characterisation of new bacteria in anaerobic digestion reactors (14, 15). On the one hand being less diverse, on the other hand being better characterised, anaerobic digestion microbiome databases are quite complete now and *a priori*, future metatranscriptomics studies might rely on these existing information, without the need of going for additional metagenome sequencing at each time. On the contrary, there are only two studies, including our own work (*Chapter 4*), reporting on the reconstruction of bacterial metagenome assembled genomes from the termite gut metagenomics study (16). The scarcity of termite gut bacterial genetic signatures in public databases, makes it in this case compulsory to combine functional characterisation of a bacterial community (metatranscriptomics) with its genomic characterisation. Especially, we showed that proper taxonomic assignment of *de novo* reconstructed gene transcripts was not possible when only relying on public databases (17) (*Chapter 4*). Moreover, combining both datasets, the presence of polysaccharide utilization loci-like clusters in re-constructed *Spirochaetes* genomes was not only evidenced here (*Chapter 4*), but I could underline the ones that were also expressed under our experimental conditions, thus prioritising them for further studies. Disintegrating both approaches would not allow for such an outcome. Another, more practical aspect of combining both approaches, that was however crucial for the success of this PhD project, was the fact that due to the applied methodology (current state-of-the-art), full length open reading frames, necessary to clone genes for further expression and heterologous production of hydrolytic enzymes, could only be retrieved from the metagenomics reconstruction. Therefore, even if the carbohydrate active enzyme candidates were identified based on their increased gene expression profiles (from metatranscriptomics datasets), only those genes that had the corresponding full-length open reading frames reconstructed in the metagenome, could have their hydrolytic capacities further assessed through heterologous protein production.

Even though clear advantages of combining both metagenomics and metatranscriptomics arose from my study, integration of higher-level approaches, including metaproteomics (protein level) and even metabolomics (metabolites level), would definitely be beneficial to further characterise these communities and their lignocellulolytic potential. Moreover, even though the analysis pipelines utilised for the purpose of my study represent the current state-of-the-art, this field of research is continuously improving, both in terms of sequencing quality and throughput, as well as of bioinformatics data treatment. Therefore, future integration of the developing long-read sequencing technologies (18), should be considered in order to generate larger and better quality datasets.

2. Despite distinct lignocellulolytic communities, the bacterial communities from the termite gut and anaerobic digestion system show similarities at the level of the carbohydrate hydrolytic potential

Lignocellulose deconstruction in the anaerobic digestion process is mainly achieved by *Bacteroidetes*, *Firmicutes* and *Chloroflexi*.

In living organisms, enzymatic hydrolysis of lignocellulose is mainly driven by carbohydrate active enzymes (19). According to the literature (e.g. (20–23)), and as shown here, the main bacterial actors of lignocellulose deconstruction in anaerobic digesters are *Bacteroidetes*, *Firmicutes*, *Chloroflexi* and to lower extent *Proteobacteria*, *Spirochaetes*, *Actinobacteria* and others (*Chapters 2 and 3*). Under stable operational conditions, in a lab-scale continuously stirred tank bio-reactor fed with sugar beet pulp, the taxonomic distribution of genes assigned to glycosyl hydrolases was different (Figure 1C *Chapter 3*) than in the case of a similar reactor facing an acidosis (Figure 1B *Chapter 2*). Still, relative metagenomic abundance of the different GH families was quite similar, with GH2, GH3, GH13, and GH43 being among the most dominant families in both cases. This observation confirms the redundancy of functions in complex bacterial communities, and aligns with the premise « function first, taxa second », recently proposed in the context of the human gut (13). On the top of this, I also suggested the retention of the hydrolytic potential

by the community not only along the stage of a stable reactor operation (*Chapter 3*), but as well during the process perturbation, here exemplified for the volatile fatty acid intoxication (*Chapter 2*). This is an interesting outcome of my thesis, as hydrolysis is often considered a bottleneck of the process. With this result I could highlight the readiness of the community that is maintained during the process failure, what allows for a quick process restauration (at least in terms of its hydrolytic capacity) once the conditions recede to favorable ones.

Lignocellulose deconstruction in termite gut systems is mainly achieved by *Spirochaetes* and *Fibrobacteres*, together with the collaboration of the host itself. In contrast to anaerobic digestion, the termite gut microbiome of the grass-feeding *Cortaritermes* sp. investigated here, was rather dominated by *Spirochaetes* and *Fibrobacteres*, and most of the assigned carbohydrate active enzymes were of that phylogenetic origin as well (*Chapter 4*). Still, other soil feeding higher termite gut microbiomes can harbor abundant *Firmicutes*, and to lower extent *Bacteroidetes* communities (17), thus somehow resembling the anaerobic digestion microbiome. Even though the genome comparison of bacteria from the two environments was not the scope of this study, a quick look at their hydrolytic capacities, based on the diversity of encoded glycosyl hydrolases, would indicate different hydrolytic potential of bacteria assigned to the same phylogenetic grouping. For example, *Firmicutes* present in anaerobic digestion reactors encode in their genomes higher number of genes assigned to GH13, GH43, GH2 and GH5 (decreasing order, *Chapter 3* Figure 4). According to Marynowska et al., *Firmicutes* in the termite gut encode mainly GH5, GH11 and GH3 (17). Similar observation was done for *Spirochaetes*. This indicates that phylogenetically similar bacteria evolved differentially in these two distinct, but functionally similar environments.

In the case of an anaerobic digestion reactor, carbohydrate hydrolytic activity is mainly of bacterial origin; on the contrary, for the termite gut system, the contribution of the host with its own carbohydrate active enzymes should not be neglected (24). In our study, we identified 170 carbohydrate active enzymes of termite origin, and one third of these represented glycoside hydrolases. Out of the 19 different families that they represented, five were not present in the termite gut microbiome, suggesting the

complementarity of hydrolytic activities expressed by the host and its gut microbiome (*Chapter 4*). Even though, the highest transcriptional abundance was attributed to host alpha glucanases (targeting alpha glucans e.g. starch), utilization of cellulose could have been deduced from the presence of abundant endoglucanases and putative lytic polysaccharide monooxygenase encoding gene transcripts from the host. What is also interesting, is the presence of a third, fungal, component in the termite gut system (other than the fungi that live in symbiosis in termite nests of fungus-growing higher termites, (25)), that was suggested based on the discovery of an auxiliary activity family 3 cellobiose oxidoreductase transcript of anaerobic gut fungi origin. In fact, with this study, we were the first one to suggest the contribution of anaerobic gut fungi to lignocellulose deconstruction in the higher termite gut system (*Chapter 4*). The diversity of termite gut fungi is currently investigated at taxonomic level by other colleagues (personal communication with Lucia Zifcakova from Okinawa Institute of Science and Technology). Possible contribution of fungi to the hydrolytic capacities of anaerobic digestion microbiome has not been critically evaluated yet. However, a recent study indicated an increased methane yield for a reactor bioaugmented with the anaerobic fungi *Orpinomycetes* (26). In my study, I did not focus on the analysis of fungi in anaerobic digestion reactors, but I suggest this topic to be further investigated.

Lignocellulose deconstruction mechanisms of bacterial origin in the two studied systems shows similarities by the presence of clusters of colocalized carbohydrate active enzymes.

In bacteria, carbohydrate active enzymes are encoded separately in the genome; they can also form clusters of co-expressed genes, e.g. cellulosomes in *Firmicutes* or polysaccharide utilization loci in *Bacteroidetes* (27, 28). While cellulosomes have been given large scientific attention in the past (29), polysaccharide utilisation loci were much less investigated. They are defined as clusters of closely located genes coding for carbohydrate active enzymes and transporters (*sus*-like genes), targeting the deconstruction of a specific polysaccharide substrate (28). Following metatranscriptomics analysis (*Chapter 3*), I was able to notice and confirm the co-expression of carbohydrate active enzymes together with their transporters (*sus*-like genes) encoded in single polysaccharide utilization loci. The first polysaccharide utilization locus was identified from *Bacteroides thetaiotaomicron*, isolated from the

human gut, and dedicated to starch deconstruction [18]. Several other polysaccharide utilization loci, specifically targeting e.g. xylan, pectins [19], xyloglucan [20], galactomannan [21], have recently been characterized. With my research, I also contributed to further characterize polysaccharide utilization loci, for the first time, proposing a mode of action for a polysaccharide utilization locus targeting an acetylated glucomannan (*Chapter 2,* Figure 5B).

Dominance of *Bacteroidetes* in the studied anaerobic digestion reactors corresponded well with the presence of multiple polysaccharide utilization loci that were identified (*Chapters 3 and 4*). They included, but are not limited to, xylan, starch as well as (acetylated)glucomannan targeting polysaccharide utilization locus (*Chapter 2*) together with (glucurono)arabinoxylan and pectin ones (*Chapter 3*). Therefore, as previously suggested (30), and as evidenced from my study, the diversity of targeted substrates, resulting from the presence of diverse polysaccharide utilization loci, might be one the reason for the abundance of *Bacteroidetes* in diverse biomass deconstructing environments.

Although widely abundant in anaerobic digestion microbiomes, *Bacteroidetes* are far less represented in the termite gut (17). Higher expression of susC and susD genes coincided with higher abundance of *Bacteroidetes* in some soil feeding higher termites, and could indicate the presence of actively transcribed polysaccharide utilization loci in the termite gut system as well (17). Moreover, according to our results, over one third of annotated carbohydrate active enzymes in the *Cortaritermes* gut microbiome were aggregated into gene clusters, containing at least two carbohydrate active enzymes, further pointing to the existence of polysaccharide utilization loci -like complexes. Indeed, similar to polysaccharide utilization loci, complexes of co-clustered carbohydrate active enzymes, however devoid of the susC/susD genes, were recently discovered in different bacteria in the termite gut (31), including *Spirochaetes* polysaccharide utilization gene clusters discovered in our study (*Chapter 5*).

Even though the two investigated systems can be classified as biomass-rich environments, bacterial communities that they harbor are phylogenetically distinct, and the hydrolytic potential of the bacteria assigned to similar taxonomic groupings

also seems different. Nevertheless, taken the diversity of termites but also the variety of anaerobic reactor designs and operation modes, the results presented here should be considered rather as preliminary, concluded based on a limited number of observations (e.g. (15, 32)). It has been suggested that each termite species operates with its own gut microbiome thus representing an endless source of new carbohydrate active enzymes (17). The scope of this study was to characterize the carbohydrate hydrolytic potential of bacteria present in the two systems with a long-term objective of their bioprospecting. Therefore, further comparison of the two systems will not be discussed here. However, one should have in mind that especially the termite gut system is much more than just its gut microbiome. Therefore, should the two systems be reliably compared, other factors should be taken into account, e.g. the presence of very distinct microhabitats along the termite gut, the influence of food chewing by the host, fungal pre-treatment in the case of some termite species, or on the other hand, mechanical mixing in the anaerobic reactors.

3. Accessory enzymes are also interesting to design nature-inspired cocktails dedicated to lignocellulose deconstruction; such cocktails are of interest to define deconstruction strategies to be exploited in the biorefinery sector

Biochemical validation for enzyme discovery and establishment of enzymatic cocktails is essential for further exploitation of newly discovered proteins in biorefinery.

In natural systems, lignocellulose deconstruction is mainly mediated by the combined action of diverse cellulases, hemicellulases and other accessory enzymes (*Chapter 1*). Therefore, the scientific community makes a huge effort to search for new lignocellulose targeting enzymes, in order to reduce the cost associated with enzymatic deconstruction by currently applied enzymatic solutions (33). Out of the different cellulases that I characterized during my PhD study, the one representing GH5_4 family and isolated from termite gut *Spirochaetae*, turned to be a multiactive enzyme, simultaneously targeting cellulose and xylan. Interestingly, deconstruction

of xylan was not expected based on the *in silico* prediction (*Chapter 5*). This finding supports the importance of a biochemical validation of the enzyme activity, in order to complement *in silico* prediction. Interestingly, the activity of the newly discovered multi-functional enzyme against carboxymethylcellulose was higher than the endocellulase from GH5 family isolated from our anaerobic digestion microbiome (*Chapter 2*). Therefore, in this special case, a reduction of the selectivity (due to its multifunction) does not seems to impact or lower the activity of the enzyme compared to monofunctional enzymes. Additionally, in comparison to single activity enzymes (34), proteins showing multiactivity are promising tools for lignocellulosic biomass deconstruction. Indeed, using proteins active against diverse substrates might be more convenient for the elaboration of cocktails compared to enzymes active against a unique substrate, i.e. optimization of enzymatic cocktail solutions is easier when less enzymes are involved. Moreover; it will allow the reduction of costs associated with proteins production, purification and activity trials.

Enzymatic pre-treatments by nature inspired cocktails might be an efficient and eco-friendly alternative in comparison to current chemical pre-treatments of lignocellulose.

Respective activities of recombinant proteins characterized during my PhD study were confirmed individually (*Chapters 3 and 5*); however one should keep in mind that enzymes act synergistically when combined in cocktails (35). As one example of application, common substrates for pulp and paper production comprises soft woods which contain acetylated glucomannan (36). Current chemical pre-treatment of these biomasses is considered as one of the most polluting activities in an industrial sector (37). Therefore, alternative pretreatment methods should be investigated (*Chapter 1*). The use of nature-evolved enzymatic cocktails specific to acetylated glucomannan deconstruction could represent such an alternative. Indeed, I could evidence that an enzymatic cocktail composed of an esterase, one exo-1,4-β-xylosidase and one β-mannanase isolated from a polysaccharide utilization locus targeting acetylated glucomannan was active against acetylated glucomannan and released monoaccharides and acetic acid (*Chapter 3*, Figure 5). Further optimization of this enzymatic cocktail was beyond the scope of my research. Therefore, these results are

only preliminary and should be further investigated by optimizing this enzymatic cocktail towards more complex substrates such as soft wood.

The use of accessory enzymes and non-hydrolytic proteins in enzymatic cocktails for pre-treatment of lignocellulosic biomass might enhance lignocellulose hydrolysis rate.

In comparison to cellulases and hemicellulases, accessory enzymes such as α-L-arabinofuranosidases and feruloyl esterases have been less investigated until now. However, they represent a high potential for the biotechnology sector as they are able to break linkages between the complex biomass components, thus increasing the efficiency of biomass deconstruction (38). Specifically, feruloyl esterases hydrolyse ester bonds between ferulic acid (which links hemicellulose to lignin) and arabinoxylan (39), while α-L-arabinofuranosidases catalyze the hydrolysis of arabinose side-chain (40), thus making the main chain of xylan more accessible to xylanases. To further investigate these two enzyme types, one of my research subjects was oriented towards their characterisation at genomic level. In both studied systems, i.e. the termite gut and the anaerobic digestion reactor, identified feruloyl esterases belonged exclusively to CE1 family, result being in line with the literature (41). Even though, genes encoding feruloyl esterases identified from the anaerobic digester and the termite gut were of different phylogenetic origins, i.e. *Bacteroidetes* and *Spirochaetes* respectively. In both cases, they were mainly encoded within arabinoxylan targeting clusters (*Spirochaetes*) or polysaccharide utilization loci (*Bacteroidetes*), together with xylanases, xylosidases and arabinofuranosidases. This observation would support the importance of accessory enzymes that assist the e.g. backbone-targeting endoxylanases in hemicellulose deconstruction (42). In this sense, enzymatic cocktails emerging from nature-optimised clusters could provide an alternative to better design and produce optimized cocktails to complement currently applied industrial treatments (43). Additionally, due to its antioxidant properties, ferulic acid resulting as by product of such enzymatic pretreatments, is also of high interest to many different sectors including food and cosmetics (41).

Although this study was mainly devoted to the characterization of carbohydrate active enzymes, both *in silico* and in heterologous systems, the addition of non-hydrolytic proteins to enzymatic cocktails might even more improve biomass deconstruction.

Indeed, it has been shown that addition of loosenin or swollenin could drastically increase the accessibility of cellulases and hemicellulases to their substrates, due to their ability to disrupt intramolecular hydrogens bonds, therefore loosening cell wall structure (33). One study reported a 300 % increase of xylose release when using swollenin together with a xylanase during corn stover deconstruction, compared to individually used enzymes (44). Additionally, Quiroz- Castaneda et al. (2011) showed an increased sugars release proportionally related to the amount of loosenin added during enzymatic pretreatment of cotton fibres with one cellulase (45). However, non-hydrolytic enzymes were not included in the scope of this PhD thesis, therefore they were not investigated in our microbiomes. Nevertheless, addition of such non-hydrolytic enzymes to standard hydrolytic cocktails could be considered in the future.

Exploring lignocellulolytic systems such as anaerobic reactors and termite guts results in a large catalogue of putatively interesting lignocellulolytic enzymes (including accessory enzymes) and opens the doors to nature-inspired cocktails for the biorefinery sector. However, one should have in mind that heterologous production of enzymes can be a fastidious process, sometimes leading to inactive proteins and resulting in low production yields (46). The methodology used for the purpose of my PhD study relied on the heterologous production of carbohydrate active enzymes in E. *coli* (as described in *Chapter 2*), a protocol shown to be time consuming, allowing for a screening of only a few proteins at the same time. Therefore, cloning techniques should be further optimized, shifting towards high-throughput cloning systems that would drastically increase the number of cloned and screened proteins. For instance, the use of protocols from commercially available cloning kits, such as Gateway (from Thermo Fisher Scientific, Waltham, MA, USA) or the Flexi Cloning system (from Promega, Madison, WI, US) was shown to be efficient tools towards high-throughput cloning (47, 48). Nevertheless, compared to more conventional cloning strategies, that are already well characterized and optimized (48), these new techniques still require huge optimization efforts related for example to the preparation of the targeted gene constructs.

4. Conclusions and perspectives

Omics-mediated characterization of microbes-driven lignocellulose deconstructing systems resulted in the discovery of new carbohydrate active enzymes. The main outputs obtained during my PhD thesis are summarized in Figure 1. Briefly, I showed that combining metagenomics and metatranscriptomics to characterize these systems holds undeniable advantages over utilizing the two approaches separately. Indeed, while metagenomics describes the **hydrolytic potential** of the microbiomes and allows the identification of bacteria with interesting capacities, here focused on lignocellulose deconstruction, metatranscriptomics allows for characterization of the **really active community members**, e.g. here based on expressed carbohydrate active enzymes. Additionally, metatranscriptomics allowed for the identification of CAZymes genes involved in lignocellulose deconstruction, which otherwise are sometimes underrepresented in the corresponding metagenomes.

Even though dominated by **phylogenetically distinct bacterial species**, i.e. *Bacteroidetes* and *Firmicutes* in the case of anaerobic digestion reactors versus *Spirochaetea* and *Fibrobacteres* in the investigated grass-feeding termite gut, the two studied systems have more similarities than previously considered. Carbohydrate metabolism seems very important to their functioning, based on the diversity and abundance of discovered CAZymes genes. Bacteria involved in lignocellulose deconstruction within the two systems harbor in their genomes **clusters of co-localised CAZymes genes**, known as **polysaccharide utilization loci** in the case of *Bacteroidetes*. Subsequent characterization of activities of recombinant proteins originating either from such gene clusters or genes encoded separately in the genomes, confirmed their classification as carbohydrate active enzymes. I demonstrated their **synergistic enzymatic activities on acetylated glucomannan**, by combining enzymes into cocktails elaborated based on their CAZyme cluster origin.

While my work helped to **identify new carbohydrate active enzymes** and to put more light on the **lignocellulose deconstruction strategies of bacteria** in anaerobic digestion reactors and the higher termite gut system, still a lot remains to be discovered. More importantly, the work done during this PhD thesis did not lead to the discovery and characterization of particularly efficient enzymes. Through my

work, I realized that efficient lignocellulose deconstruction relies on the **complementarity of various enzymes**, their sequential action, and the community contribution **rather than on single super-enzymes**.

To further **complement the results obtained** during my study and to increase our knowledge on the deconstruction of highly lignified lignocellulosic biomass, especially within **anaerobic digestion**, metagenomics and metatranscriptomics should be applied to reactors fed with substrates known to be more recalcitrant than sugar beet pulp, e.g. **miscanthus**. This strategy was proposed here for the termite gut system, by feeding and analyzing the termite gut digestome (set of CAZymes) under miscanthus diet. **Supplementary multi-omics** approach (metaproteomics and/or metabolomics) should also be applied to the studied systems. It would definitively bring more comprehensive knowledge on the importance of the bacterial community composition, the bacterial interactions, and the host-microbe interactions involved in lignocellulose deconstruction as well as the discovery of key enzymes involved in the process.

Another perspective arising from the results of my study would be the analysis of bacteria with high hydrolytic potential, though not studied in this PhD. Here, especially ***Planctomycetes*** could represent a high potential for the biorefinery sector owing to their diverse and **large CAZymes repertoire**, Nevertheless, current constrains related to our ability to cultivate new microorganisms, would have to be addressed first.

Another extension of this PhD thesis could be the application of the techniques and methodologies used here, but focused on novel and efficient enzymes involved in the **removal of lignin** from the lignocellulose complexes and general aspects of lignin utilization. To that purpose, best lignin-utilization systems and organisms should be first identified. Such approach would help clarifying if sequential enzymatic pre-treatment of lignocellulosic biomass can be an alternative to obtain pure lignocellulose components with high added value for the biorefenery sector, such as lignin.

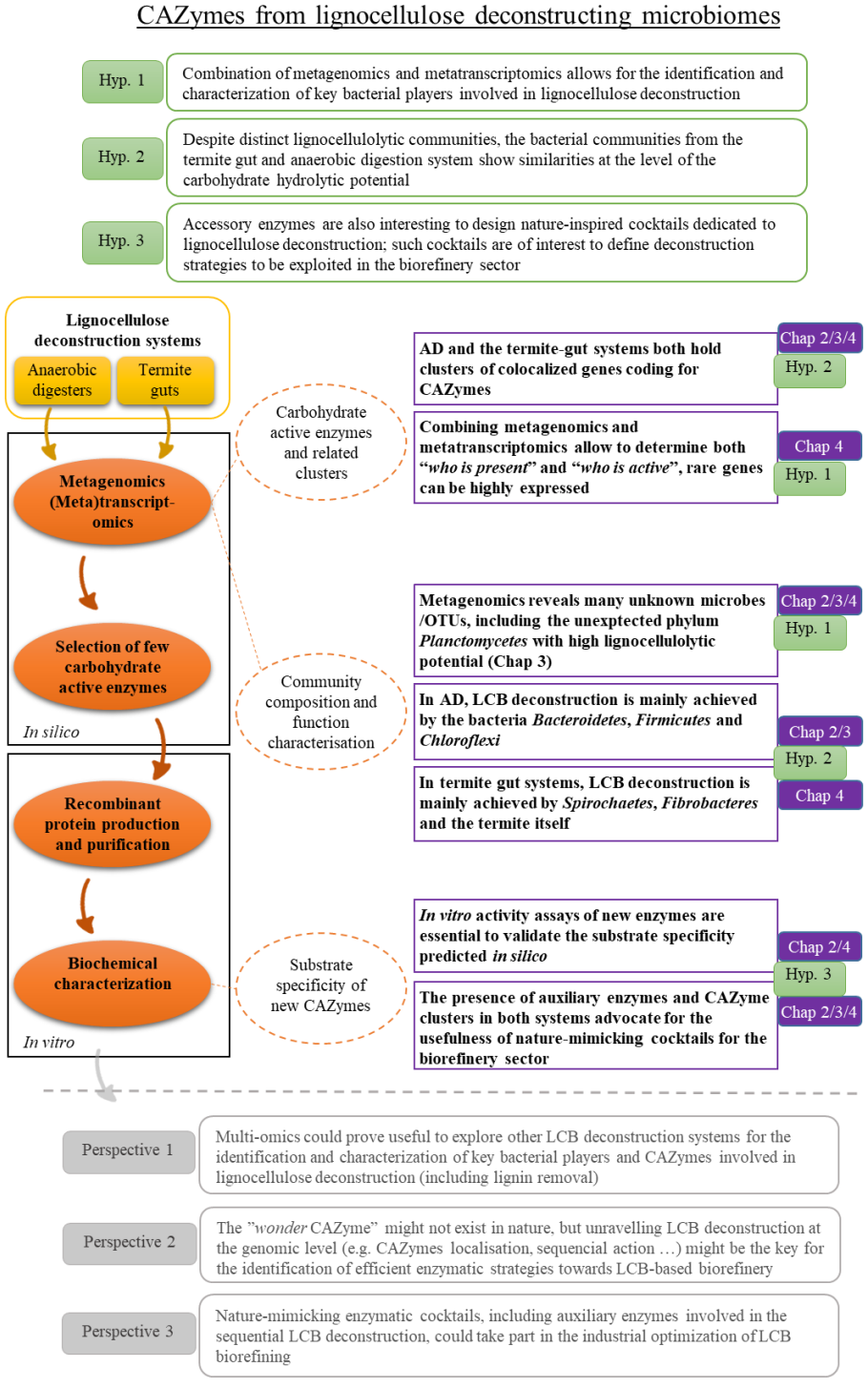

CAZymes from lignocellulose deconstructing microbiomes
Hyp. 1
Combination of metagenomics and metatranscriptomics allows for the identification and characterization of key bacterial players involved in lignocellulose deconstruction
Hyp. 2
Despite distinct lignocellulolytic communities, the bacterial communities from the termite gut and anaerobic digestion system show similarities at the level of the carbohydrate hydrolytic potential
Hyp. 3
Accessory enzymes are also interesting to design nature-inspired cocktails dedicated to lignocellulose deconstruction; such cocktails are of interest to define deconstruction strategies to be exploited in the biorefinery sector
Lignocellulose deconstruction systems
Anaerobic digesters
Termite guts
Metagenomics (Meta)transcript-omics
Selection of few carbohydrate active enzymes
In silico
Recombinant protein production and purification
Biochemical characterization
In vitro
Carbohydrate active enzymes and related clusters
Community composition and function characterisation
Substrate specificity of new CAZymes
AD and the termite-gut systems both hold clusters of colocalized genes coding for CAZymes
Chap 2/3/4
Hyp. 2
Combining metagenomics and metatranscriptomics allow to determine both "who is present" and "who is active", rare genes can be highly expressed
Chap 4
Hyp. 1
Metagenomics reveals many unknown microbes /OTUs, including the unexptected phylum Planctomycetes with high lignocellulolytic potential (Chap 3)
Chap 2/3/4
Hyp. 1
In AD, LCB deconstruction is mainly achieved by the bacteria Bacteroidetes, Firmicutes and Chloroflexi
Chap 2/3
Hyp. 2
In termite gut systems, LCB deconstruction is mainly achieved by Spirochaetes, Fibrobacteres and the termite itself
Chap 4
In vitro activity assays of new enzymes are essential to validate the substrate specificity predicted in silico
Chap 2/4
Hyp. 3
The presence of auxiliary enzymes and CAZyme clusters in both systems advocate for the usefulness of nature-mimicking cocktails for the biorefinery sector
Chap 2/3/4
Perspective 1
Multi-omics could prove useful to explore other LCB deconstruction systems for the identification and characterization of key bacterial players and CAZymes involved in lignocellulose deconstruction (including lignin removal)
Perspective 2
The "wonder CAZyme" might not exist in nature, but unravelling LCB deconstruction at the genomic level (e.g. CAZymes localisation, sequencial action …) might be the key for the identification of efficient enzymatic strategies towards LCB-based biorefinery
Perspective 3
Nature-mimicking enzymatic cocktails, including auxiliary enzymes involved in the sequential LCB deconstruction, could take part in the industrial optimization of LCB biorefining

Figure 1: CAZymes from lignocellulose deconstructing microbiomes. Schematic representation of the approach undertaken in the course of this study. Investigation of two different lignocellulose deconstruction systems by combining *in silico* analysis (metagenomics and metatranscriptomics) and *in vitro* analysis (recombinant protein production and biochemical characterization). The different hypotheses discussed in the frame of this PhD thesis are also represented and are linked to the related outputs (framed in purple). Some perspective arising from this PhD thesis are presented. LCB refers to lignocellulosic biomass, CAZymes refers to carbohydrate active enzymes, Hyp. refers to hypothesis. Chap refers to chapter. AD refers to anaerobic digester. OTUs refers to operational taxonomic unit

References

1. B. S. Boneberg, *et al.*, Biorefinery of lignocellulosic biopolymers. *Rev. Eletrônica Científica da UERGS* **2**, 79 (2016).
2. L. J. Jönsson, C. Martín, Pretreatment of lignocellulose: Formation of inhibitory by-products and strategies for minimizing their effects. *Bioresour. Technol.* **199**, 103–112 (2016).
3. S. Y. Lee, *et al.*, Waste to bioenergy: a review on the recent conversion technologies. *BMC Energy* **1**, 1–22 (2019).
4. H. Luo, *et al.*, Total Utilization of Miscanthus Biomass, Lignin and Carbohydrates, Using Earth Abundant Nickel Catalyst. *ACS Sustain. Chem. Eng.* **4**, 2316–2322 (2016).
5. A. Bayané, S. R. Guiot, Animal digestive strategies versus anaerobic digestion bioprocesses for biogas production from lignocellulosic biomass. *Rev. Environ. Sci. Biotechnol.* **10**, 43–62 (2011).
6. A. Brune, Symbiotic digestion of lignocellulose in termite guts. *Nat. Rev. Microbiol.* **12**, 168–180 (2014).
7. C. Sawatdeenarunat, K. C. Surendra, D. Takara, H. Oechsner, S. K. Khanal, Anaerobic digestion of lignocellulosic biomass: Challenges and opportunities. *Bioresour. Technol.* **178**, 178–186 (2015).
8. M. Hess, *et al.*, Metagenomic Discovery of Biomass-Degrading Genes and genomes from Cow Rumen. *Science (80-.).* **463**, 463–467 (2011).
9. R. Berlemont, A. C. Martiny, Genomic Potential for Polysaccharide Deconstruction in Bacteria. *Appl. Environ. Microbiol.* **81**, 1513–1519 (2015).
10. A. Tartar, *et al.*, Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite Reticulitermes flavipes. *Biotechnol. Biofuels* **2**, 1–19 (2009).
11. R. Raychoudhury, *et al.*, Comparative metatranscriptomic signatures of wood and paper feeding in the gut of the termite Reticulitermes flavipes (Isoptera: Rhinotermitidae). *Insect Mol. Biol.* **22**, 155–171 (2013).
12. S. He, *et al.*, Comparative Metagenomic and Metatranscriptomic Analysis of Hindgut Paunch Microbiota in Wood- and Dung-Feeding Higher Termites. *PLoS One* **8** (2013).
13. A. Heintz-Buschart, P. Wilmes, Human Gut Microbiome: Function Matters. *Trends Microbiol.* **26**, 563–574 (2018).
14. S. Campanaro, *et al.*, The anaerobic digestion microbiome: a collection of 1600 metagenome-assembled genomes shows high species diversity related to methane

production. *bioRxiv*, 680553 (2019).
15. S. Campanaro, L. Treu, P. G. Kougias, G. Luo, I. Angelidaki, Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water Res.* **140**, 123–134 (2018).
16. V. Hervé, *et al.*, *Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites* (2019) https:/doi.org/10.7287/peerj.preprints.27929v1.
17. M. Marynowska, *et al.*, Compositional and functional characterisation of biomass-degrading microbial communities in guts of plant fibre- And soil-feeding higher termites. *Microbiome* **8**, 1–18 (2020).
18. S. L. Amarasinghe, *et al.*, Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 1–16 (2020).
19. C. Álvarez, F. M. Reyes-Sosa, B. Díez, Enzymatic hydrolysis of biomass from wood. *Microb. Biotechnol.* **9**, 149–156 (2016).
20. L. Sun, P. B. Pope, V. G. H. Eijsink, A. Schnürer, Characterization of microbial community structure during continuous anaerobic digestion of straw and cow manure. *Microb. Biotechnol.* **8**, 815–827 (2015).
21. X. Goux, *et al.*, Microbial community dynamics in replicate anaerobic digesters exposed sequentially to increasing organic loading rate, acidosis, and process recovery. *Biotechnol. Biofuels* **8** (2015).
22. M. Zamanzadeh, L. H. Hagen, K. Svensson, R. Linjordet, S. J. Horn, Anaerobic digestion of food waste - Effect of recirculation and temperature on performance and microbiology. *Water Res.* **96**, 246–254 (2016).
23. D. Rivière, *et al.*, Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *ISME J.* **3**, 700–714 (2009).
24. M. Poulsen, *et al.*, Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 14500–14505 (2014).
25. S. Otani, *et al.*, Identifying the core microbial community in the gut of fungus-growing termites. *Mol. Ecol.* **23**, 4631–4644 (2014).
26. Ç. Akyol, O. Ince, M. Bozan, E. G. Ozbayram, B. Ince, Fungal bioaugmentation of anaerobic digesters fed with lignocellulosic biomass: What to expect from anaerobic fungus Orpinomyces sp. *Bioresour. Technol.* **277**, 1–10 (2019).
27. L. Artzi, E. A. Bayer, S. Moraïs, Cellulosomes: Bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15**, 83–95 (2017).
28. J. M. Grondin, K. Tamura, G. Déjean, D. W. Abbott, H. Brumer, Polysaccharide Utilization Loci: Fuelling microbial communities. *J. Bacteriol.* **199**, JB.00860-16 (2017).
29. B. Yang, Z. Dai, S. Y. Ding, C. E. Wyman, Enzymatic hydrolysis of cellulosic biomass. *Biofuels* **2**, 421–450 (2011).
30. F. Thomas, J. H. Hehemann, E. Rebuffet, M. Czjzek, G. Michel, Environmental and gut Bacteroidetes: The food connection. *Front. Microbiol.* **2**, 1–16 (2011).
31. N. Liu, *et al.*, Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *ISME J.*, 104–117 (2019).
32. M. Calusinska, *et al.*, A year of monitoring 20 mesophilic full - scale bioreactors reveals the existence of stable but different core microbiomes in bio - waste and wastewater anaerobic digestion systems. *Biotechnol. Biofuels* **11** (2018).
33. A. M. Lopes, E. X. Ferreira Filho, L. R. S. Moreira, An update on enzymatic

cocktails for lignocellulose breakdown. *J. Appl. Microbiol.* **125**, 632–645 (2018).
34. D. Talamantes, N. Biabini, H. Dang, K. Abdoun, R. Berlemont, Natural diversity of cellulases, xylanases, and chitinases in bacteria. *Biotechnol. Biofuels* **9**, 1–11 (2016).
35. S. Malgas, R. Chandra, J. S. Van Dyk, J. N. Saddler, B. I. Pletschke, Formulation of an optimized synergistic enzyme cocktail, HoloMix, for effective degradation of various pre-treated hardwoods. *Bioresour. Technol.* **245**, 52–65 (2017).
36. P. M.-A. Pawar, S. Koutaniemi, M. Tenkanen, E. J. Mellerowicz, Acetylation of woody lignocellulose: significance and regulation. *Front. Plant Sci.* **4**, 118 (2013).
37. C. Veluchamy, A. S. Kalamdhad, Influence of pretreatment techniques on anaerobic digestion of pulp and paper mill sludge: A review. *Bioresour. Technol.* **245**, 1206–1219 (2017).
38. G. Banerjee, J. S. Scott-Craig, J. D. Walton, Improving enzymes for biomass conversion: A basic research perspective. *Bioenergy Res.* **3**, 82–92 (2010).
39. D. M. Oliveira, *et al.*, Feruloyl esterases: Biocatalysts to overcome biomass recalcitrance and for the production of bioactive compounds. *Bioresour. Technol.* **278**, 408–423 (2019).
40. S. Lagaert, A. Pollet, C. M. Courtin, G. Volckaert, β -Xylosidases and α - L - arabinofuranosidases : Accessory enzymes for arabinoxylan degradation. *Biotechnol. Adv.* **32**, 316–332 (2014).
41. A. Dilokpimol, *et al.*, Diversity of fungal feruloyl esterases : updated phylogenetic classification , properties , and industrial applications. *Biotechnol. Biofuels* **9**, 1–18 (2016).
42. D. Xin, X. Chen, P. Wen, J. Zhang, Insight into the role of α-arabinofuranosidase in biomass hydrolysis: cellulose digestibility and inhibition by xylooligomers. *Biotechnol. Biofuels* **12**, 1–11 (2019).
43. A. E. Fazary, Y. H. Ju, Feruloyl esterases as biotechnological tools: Current and future perspectives. *Acta Biochim. Biophys. Sin. (Shanghai).* **39**, 811–828 (2007).
44. K. Gourlay, *et al.*, Swollenin aids in the amorphogenesis step during the enzymatic hydrolysis of pretreated biomass. *Bioresour. Technol.* **142**, 498–503 (2013).
45. R. E. Quiroz-Castañeda, C. Martínez-Anaya, L. I. Cuervo-Soto, L. Segovia, J. L. Folch-Mallol, Loosenin, a novel protein with cellulose-disrupting activity from Bjerkandera adusta. *Microb. Cell Fact.* **10**, 8 (2011).
46. G. L. Rosano, E. a. Ceccarelli, Recombinant protein expression in Escherichia coli: Advances and challenges. *Front. Microbiol.* **5**, 1–17 (2014).
47. T. Nagase, *et al.*, Exploration of human ORFeome: High-throughput preparation of ORF clones and efficient characterization of their protein products. *DNA Res.* **15**, 137–149 (2008).
48. B. Jia, C. O. Jeon, High-throughput recombinant protein expression in Escherichia coli: Current status and future perspectives. *Open Biol.* **6** (2016).

Appendices

Annex 1: Protocols optimization for recombinant protein production

1. Selection, amplification and initial cloning in pGem-t-easy of a gene of interest

Following -OMICS analysis, genes of interest (GOI) were identified. The presence of any signal peptide on the protein was predicted online via LipoP version 1.0 (1). Specific primers for each GOI were designed in *silico*, according to their respective DNA sequence (Figure 1). Briefly, start and stop codons were removed. Forward primer corresponds to the 5' end of the gene sequence (if any putative signal peptide was predicted and if specified, it was *in silico* removed prior to the design of the forward primer), however, reverse primer correspond to the reverse complement of the 3' end. Primer sequences of different length were submitted to an online tool (http://tmcalculator.neb.com) to determine a pair of primers showing identical or very similar specifications, i.e. annealing temperature. Once good candidates were identified, addition of restriction sites (usually BamHI for the forward primer and SacI for the reverse primer, unless specified differently) was performed *in silico*. Primers were synthetised by Eurogentec (Eurogentec, Belgium). Finally, the GOI was amplified using the total DNA extracted from the studied sample by polymerase chain reaction (PCR), and it was further cloned to the intermediate vector pGEM®-T-Easy. This two-step cloning was more efficient than a single cloning protocol involving a direct directional cloning to an expression verctor. Please refer to the material and methods section of chapters 2 and 4 for more details.

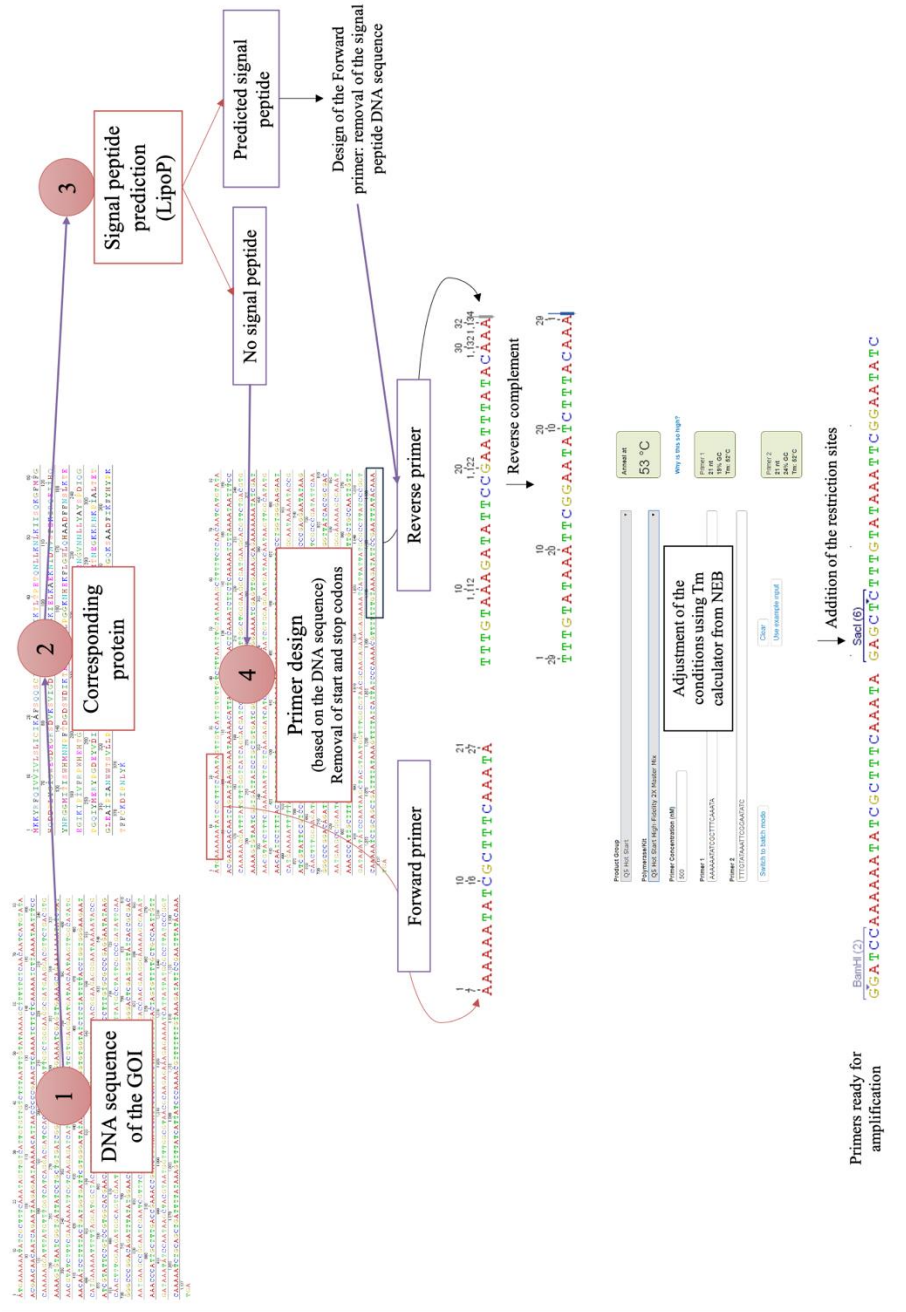


Figure 1: Pipeline used to design specific primers for the amplification of the GOI. Signal peptide (SP) prediction was performed online via LipoP version 1.0 (1). Specific primers of the GOI were designed based on its DNA sequence (if SP predicted and if specified, forward primer was designed in order to remove the SP)

2. Heterologous protein production in *E. coli*

With the use of specific restriction enzymes, GOI was re-cloned into an expression vector and initially expressed in *E. coli*. Due to its IPTG-inducible promoter, flexible cloning site (which includes multiple restriction sites) and its dual tag, pET52b(+) (Figure 2) was chosen as best plasmid candidate. This vector includes a streptavidin tag (Strep-tag), that is located at the N-terminus of the protein, as well as a histidine tag (His tag) at the C-terminus, both were used during further steps of protein purification. They consist of specific peptide sequences allowing a detection of the recombinant protein and its purification by affinity chromatography. Strep-tag shows a high affinity to streptavidin while His-tag exhibits a high affinity with divalent nickel or cobalt ions (2, 3). Affinity-tag purification is a commonly used method in recombinant protein production experiments (4–6).

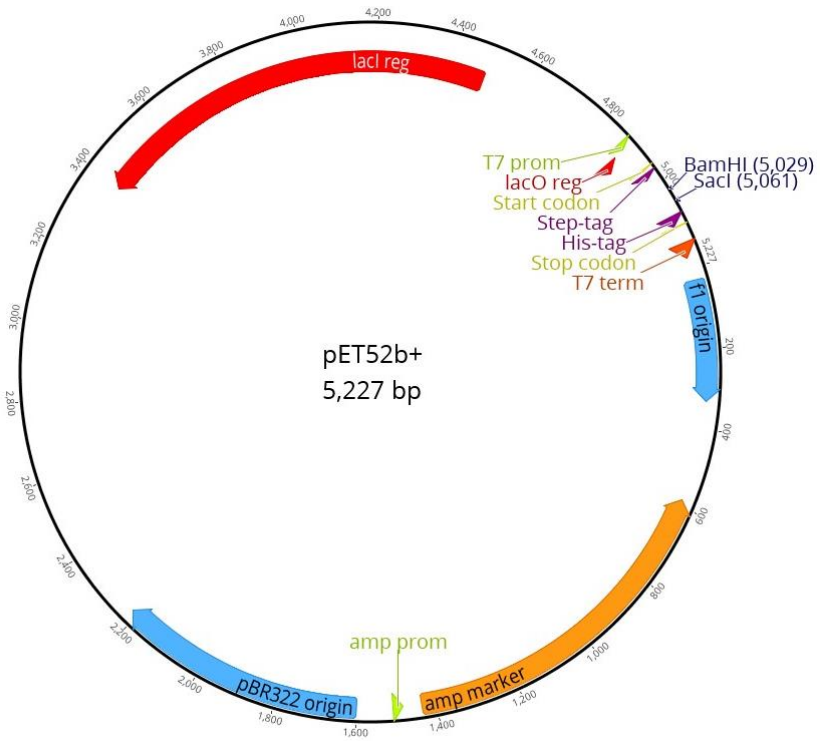


Figure 2: The pET52b(+) vector used to clone the genes of interest in *E. coli.* This vector is IPTG inducible and contains a Strep-tag located at the N-terminus of the recombinant protein as well as a His-tag located at the C-terminus of the recombinant protein.

A specifically adapted version of the protocol is given in each chapter that deals with recombinant protein production. The cloning vector (pGEM®-T-Easy) containing the GOI, and the expression vector (pET52b(+)) were digested with the appropriate restriction enzymes (purchased from New England Biolabs, USA), following the supplier's protocol. Briefly, 2 µg of expression vector or 5 µg of the cloning vector containing the GOI was incubated with 5 µL of Cut Smart buffer, 1.5 µL of SacI and water was added to reach a total volume of 48.5 µL. The solution was incubated at 37°C for 30 minutes and then the reaction was stopped by heating at 65°C for 20 min. Finally, 1.5 µL of BamHI was added to the reaction medium that was incubated at 37°C for additional 30 minutes. The reaction was stopped by adding 10 µL of 6X loading dye and purification of the GOI was performed from an agarose gel using a commercially available kit. After that, the GOI was ligated into the opened vector and introduced in *E. coli* cells by heat shock. Glycerol stocks of transformed cells were stored at -80°C and used whenever production of the protein was needed. Protein expression was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG). Details of the production procedure can be found in the material and methods section in chapters 2 and 4. Production of the recombinant protein was compared between two different strains of *E. coli*, BL21(DE3) and Rosetta(DE3). BL21(DE3) is a strain mainly used for CAZymes production in the literature (7–9), while Rosetta(DE3) is a mutant that provides few codons rarely used in *E. coli*. The advantage of Rosetta(DE3) cells might be a higher production of proteins that originate from bacterial species distantly related to *E. coli* (e.g. as it might be the case for *Bacteroidetes* sp.), and thus capable of expressing codons that are otherwise rare in *E. coli* (10, 11).

3. Heterologous protein production in *B. megaterium*

B. megaterium was chosen for its ability to produce recombinant proteins extracellularly. Indeed, it facilitates the continuous production of the protein as no cell lysis is needed to harvest the recombinant protein. Following the same protocol as for the production in *E. coli*, the cloning vector containing the GOI, pGEM®-T-Easy, and the expression vector pSTREPHIS1525 (MoBiTec, Germany) (Figure 3) were digested with the restriction enzymes and ligated together. The expression vector used for expression in *B. megaterium* was chosen in order to be in line with the one for the expression in E. *coli* (i.e. it also contains the two tags), and also in order to be able to produce the protein extracellularly (it harbours a signal peptide). Once the GOI was ligated into the appropriate vector, transformation and expression in *B. megaterium* protoplasts was performed according to the supplier's protocol (MoBiTec, Germany).

Ligation of the GOI in the pSTREPHIS1525 vector was quite tricky due to its large size, and despite several attempts, it was not achieved in most cases. Even though several proteins were cloned and could have been heterologously produced in *B. megaterium* protoplasts, the protein yield was comparable or even lower than in the case of *E. coli*. Therefore, due to the technical difficulties related to protein production using this system, it was not further optimised in this study. Thus, most of the characterised proteins were produced using *E. coli.*

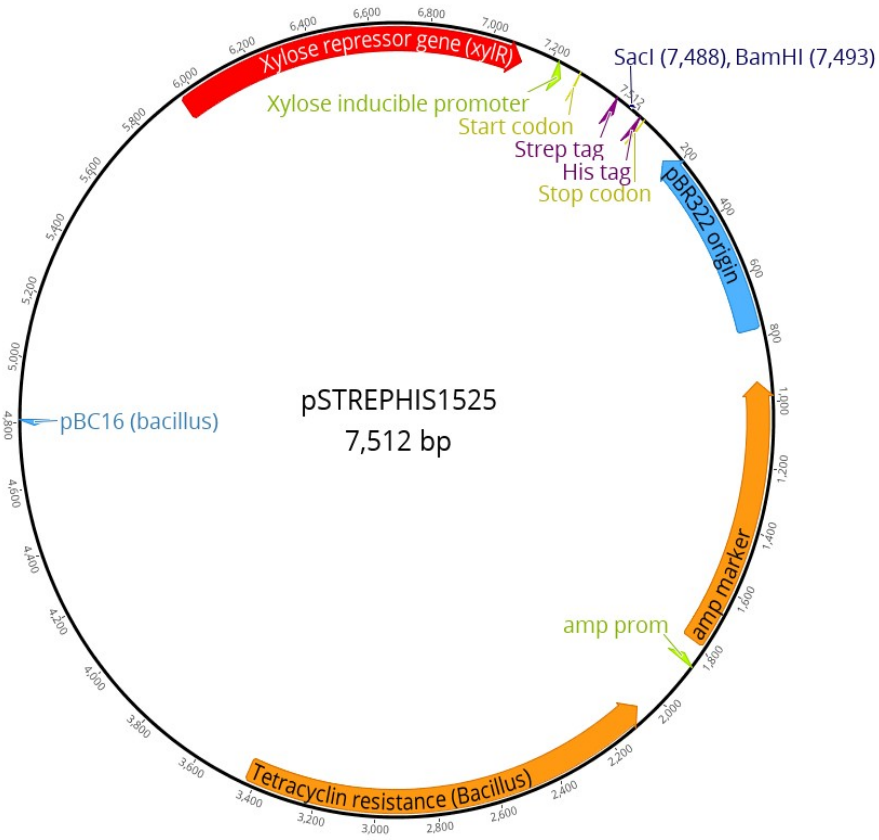


Figure 3: The pSTREPHIS1525 vector used to clone the genes of interest in *B. megaterium.* The vector used is xylose inducible and contains a signal peptide (allowing extracellular production of the recombinant protein), a Strep-tag located at the N-terminus of the recombinant protein as well as a His-tag located at the C-terminus of the recombinant protein.

4. Protein recovery

As the proteins were produced intracellularly in *E. coli*, cell lysis was essential to release the cell content. Therefore, cells were collected by centrifugation and resuspended in an appropriate lysis buffer (depending on the purification method). Lysis was performed by sonication, with a protocol adapted from previous studies (12, 13). One cycle of lysis is defined as 2-min pulse (1 s on/1s off) followed by a 2-min pause (40% of amplitude). Different number of cycles were tested in order to ensure a complete lysis of the cells. Following the lysis, removal of the cell debris was

achieved by centrifugation and filtration. Detailed protocol can be found in the material and methods section in chapters 2 and 4. Results were analysed on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) (protein separation based on their molecular weight). Figure 4 shows the comparison between the expression in the two *E. coli* strains, as well as the results of the cell lysis.

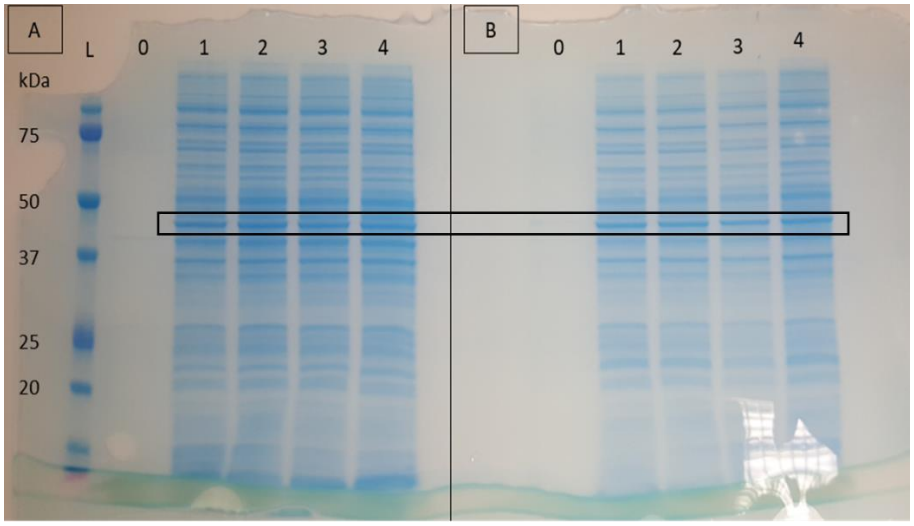


Figure 4: SDS-PAGE evidencing the impact of increased number of cell lysis sonication cycle on the release of a recombinant proteins (RP) of theorical molecular weight about 45 kDa (highlighted by a black solid frame). **A** - Expression of RP in *E. coli* BL21(DE3). **B** - Expression of RP in *E. coli* Rosetta(DE3). Lane L: Molecular mass marker.Lane 0: 0 cycle of lysis. Lane 1: 1 cycle of lysis. Lane 2: 2 cycles of lysis. Lane 3: 3 cycles of lysis. Lane 4: 4 cycles of lysis.

According to the generated results, virtually no difference in intensity (quantity of released proteins) was observed, and *E. coli* Rosetta(DE3) was selected as the recombinant host. Additionally, the number of sonication cycles necessary for a complete release of the proteins was set at 2 cycles.

Regarding proteins produced by *B. megaterium*, they were already present in the extracellular medium, therefore precipitation of the recombinant protein was achieved following the supplier's protocol (MoBiTec, Germany). Unfortunately, the expression as well as the protein precipitation was tricky, and did not lead to high amounts of recombinant proteins (data not shown). In order to get higher expression yield, it would require a lot of optimization. However, we decided to stop at this stage for the

expression in *B. megaterium* and to carry on with *E. coli* Rosetta(DE3) only, as this expression system led to sufficient yields of protein production.

5. Protein detection

The detection of the recombinant protein, using both tags, was carried out by Western blotting. Detailed protocol can be found in the material and methods section in chapter 2. The transfer of proteins was done on a nitrocellulose membrane and detection was performed by enhanced chemiluminescence. Procedure for the detection using the two tags was identical, except for the antibody used. For the His-tag detection the antibody used was a polyclonal 6x His-tag (ThermoFisher, USA), while for Strep-tag the antibody used was Strep-tactin HRP conjugate (IBA Lifesciences, Germany), both antibodies were labelled with horseradish peroxidase. The comparison of both tags shows that the detection was better for His-tag than Strep-tag (Figure 5A).

6. Protein production with gravity flow columns and NGC

Following the detection of the presence of the produced protein, its purification was performed by gravity flow chromatography columns. Similarly, comparison of the two affinity-tags was done. For Strep-tag purification, cell lysate (previously dissolved in lysis/washing buffer composed of of 100 mM Tris/HCl pH8, 150 mM NaCl, 1 mM EDTA), was loaded onto a Strep-tactin gravity flow column (IBA lifesciences, Germany). Washing and elution was performed according to the supplier's instructions (elution buffer was composed as follow, 100 mM Tris/HCl, pH8, 150 mM NaCl, 1 mM EDTA. 50 mM biotin). Purification using His-tag was performed as explained later in the material and methods section in chapters 2 and 4. Both purification procedures were compared (Figure 5B), and in line with the results on protein detection, the purification using His-tag resulted in higher quantities of partially purified proteins. Nevertheless, in some cases, proteins were truncated at the N-terminus (e.g. Figure 5), therefore His-tag purification was retained as a standard protocol for protein purification.

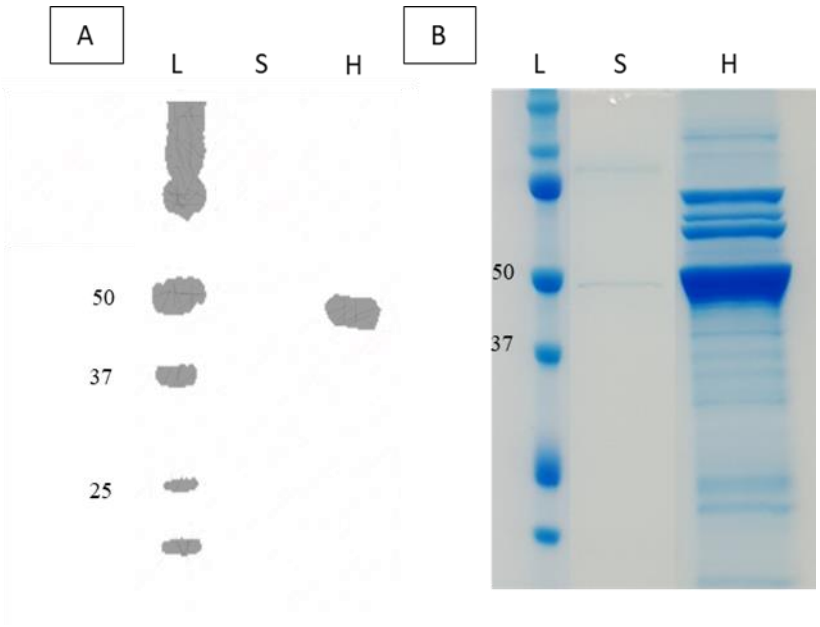


Figure 5: Comparison of the two investigated tags applied to a recombinant protein (RP) of theorical molecular weight about 50 kDa expressed in *E. coli* Rosetta(DE3). **A** - Detection of the RP using Strep-tag (lane S) and His-tag (lane H) and evidenced by western blot on a membrane revealed by enhanced chemiluminescence (oxidation of luminol by horse-radish peroxidase coupled to the anti-Strep Tag II monoclonal (lane S) and anti-6X-His polyclonal (lane H) antibodies). **B** - Purification of the RP using Strep-tag (lane S) and His-tag (lane H) evidenced by SDS-PAGE. Lane L: molecular weight marker

Depending on the purpose of the study, when needed, the recombinant protein was identified following a MALDI-TOF-TOF analysis; detailed protocol can be found in the supplementary material of chapter 2 (Annex 2).

7. Activity tests

Release and monitoring of 4-nitrophenol from the hydrolysis of 4-nitrophenyl derivative substrates

One type of substrates that we used for the detection of enzymatic activity was 4-nitrophenyl derivatives. This synthetic substrate is composed of a 4-nitrophenol molecule linked to a mono/di-sugar (Figure 6).

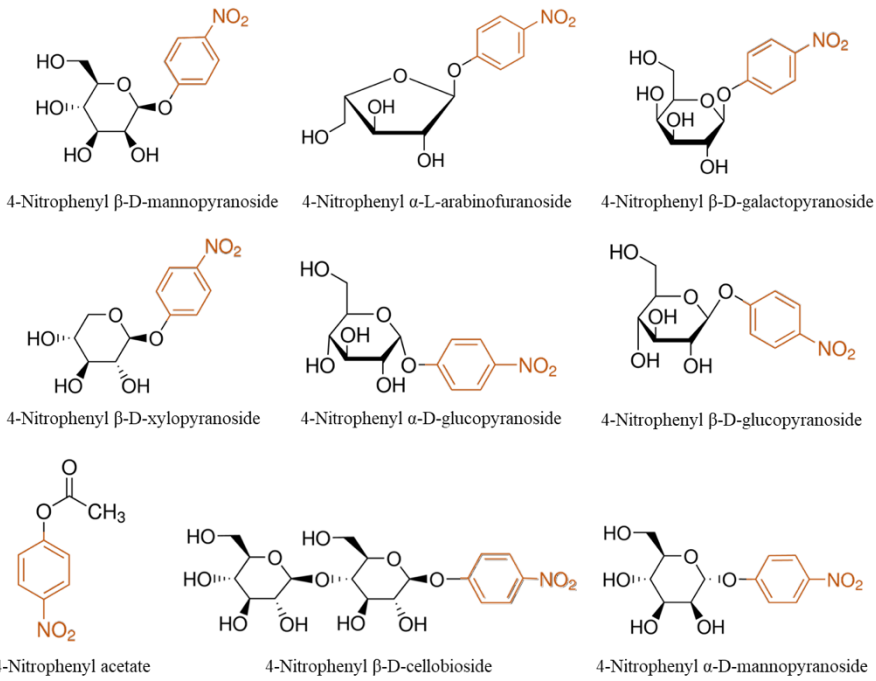


Figure 6: Representation of the synthetic substrates used for unravelling enzymatic activity in this study. The 4-nitrophenol moeity is coloured in orange.

The hydrolysis of the bond between the substrate and the 4-nitrophenol results in a yellowish coloration of the solution that can be spectrophotometrically monitored at 405 nm. Therefore, the detection of the activity is straightforward when dealing with the appropriate enzyme (more especially exo-acting CAZymes, e.g hydrolyzing glycosidic bonds at the end of the polysaccharide chain) able to hydrolyse the specific bond. Details about the procedure used to perform enzymatic assays can be found in the material and methods section of the respective chapters 2 and 4. The specific activity of the studied enzyme is defined as the amount of substrate converted per minute and per milligram of protein. Depending on the study, the relative activity of a protein might be used (the maximum activity observed for a substrate under specific conditions is set up to 100%).

Unfortunately, this type of substrate is inappropriate for certain type of enzymes, especially the endo-active CAZymes (e.g. endoxylanases), the latest targeting long chain polysaccharides as substrate. In that case, another type of assays for the activity detection needs to be done.

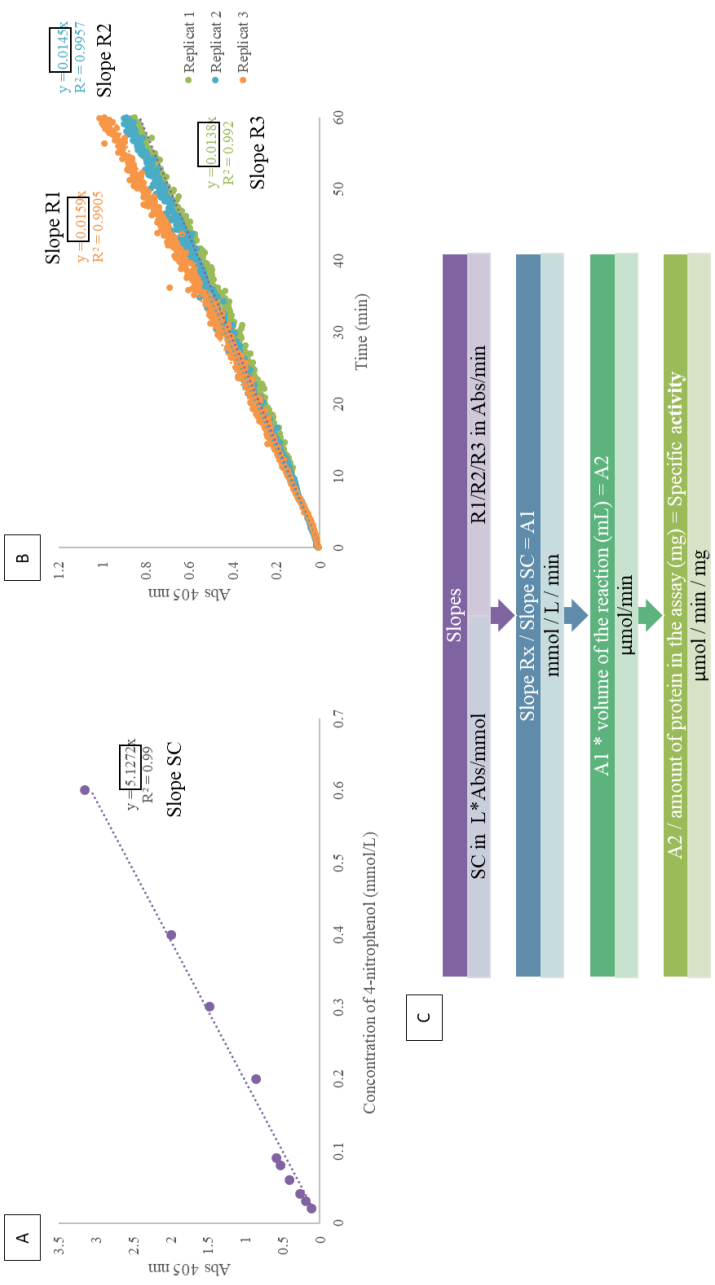


Figure 7: Workflow for the determination of the enzymatic activity using 4-nitrophenol derivatives. **A** - Calibration curve of 4-nitrophenol allowing the determination of the concentration of 4-nitrophenol against absorbance at 405 nm, in the studied enzymatic reactions. **B** - Example of monitoring of 4-nitrophenol release during enzymatic hydrolysis of the substrate. **C** - Calculations applied to the results

Release and monitoring of reducing sugars from the hydrolysis of polysaccharide substrates

Endo-acting CAZymes are hydrolyzing internal glycosidic linkages in the polysaccharide chain. One of the method that can be used to determine endo-acting activities is the determination of the release of reducing sugars using the Somogyi-Nelson method (14). This method is based on the principle that the hydrolysis of some polysaccharides (e.g. xylan, cellulose or mannan) by their respective endo-acting enzymes (e.g. xylanases, cellulases or mannanases) result in the release of reducing sugars. Interestingly, cupric ions (Cu2+) are reduced to cuprous ions (Cu+) in the presence of reducing sugars. Finally, an arsenomolibdate complex is reduced by Cu+, producing a blue coloration (proportional to the concentration of reducing sugar) that can be measured spectrophotometrically at 620 nm. For this method, the results of the enzymatic reaction were measured after 30 min. Details on the enzymatic activity measurement can be found in the material and methods section in chapters 2 and 4. Similarly to the determination of the activity based on 4-nitrophenol derivatives, activity was defined as the amount of enzyme catalyzing (in mg) the conversion of one µmol of reducing sugar per minute. Glucose was used as the standard reducing sugar (Figure 8). Depending on the study, the relative activity of a protein might be used (the maximum activity observed for a substrate under specific conditions is set up to 100%).

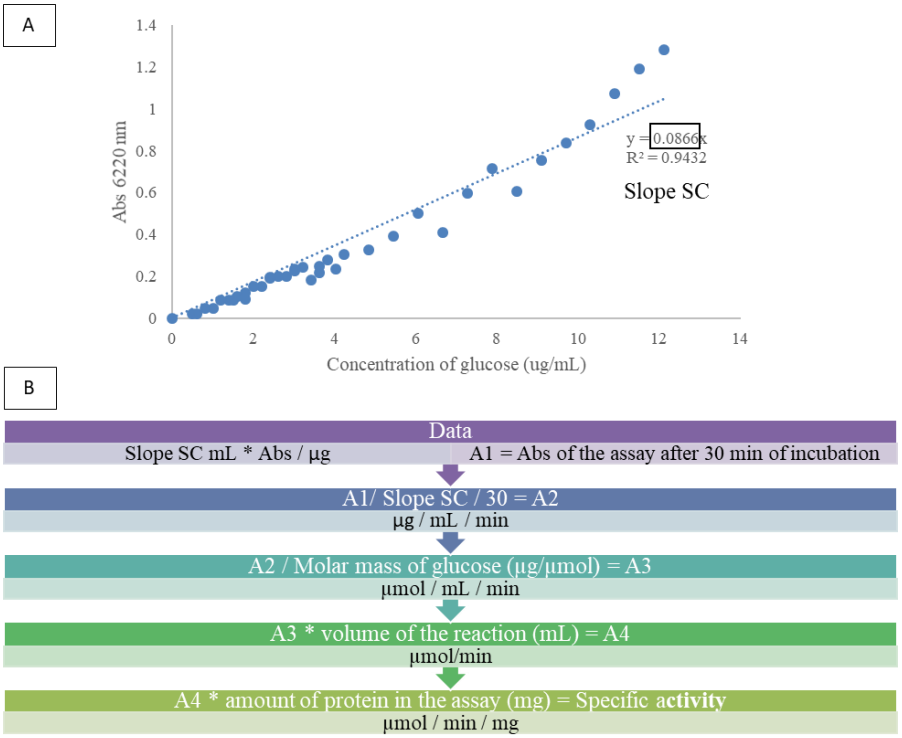


Figure 8: Workflow for the determination of the enzymatic activity using polysaccharides. **A** - Calibration curve of glucose standard allowing the determination of the concentration of reducing sugars against absorbance at 620 nm, in the studied enzymatic reactions. **B** - Calculations applied to the results

Release and monitoring of specific molecules from the hydrolysis of polysaccharide substrates

In some cases, enzymatic cocktails were tested against specific polysaccharides. The goal of this approach was to compare the release of specific molecules (e.g. glucose, xylose, acetic acid, etc. …) when different cocktails of enzymes were applied to the same substrate. Details about this procedure can be found in the material and methods section in chapter 2. The concentration of the released reducing carbohydrate was determined following commercially available kits purchased from Megazyme (Megazyme, Ireland).

References

1. A. S. Juncker, *et al.*, Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652–1662 (2003).
2. T. G. M. Schmidt, J. Koepke, R. Frank, A. Skerra, Molecular interaction between the strep-lag affinity peptide and its cognate target, streptavidin. *J. Mol. Biol.* **255**, 753–766 (1996).
3. A. Hoffmann, R. G. Roeder, Purification of his-tagged proteins in non-denaturing conditions suggests a convenient method for protein interaction studies. *Nucleic Acids Res.* **19**, 6337–6338 (1991).
4. F. Cuskin, *et al.*, The GH130 family of mannoside phosphorylases contains glycoside hydrolases that target β-1,2-mannosidic linkages in Candida mannan. *J. Biol. Chem.* **290**, 25023–25033 (2015).
5. J. Gao, W. Wakarchuk, Characterization of Five B-Glycoside Hydrolases from Cellulomonas fimi ATCC 484. *J. Bacteriol.* **196**, 4103–4110 (2014).
6. H. Zheng, *et al.*, Overexpression of a Paenibacillus campinasensis xylanase in Bacillus megaterium and its applications to biobleaching of cotton stalk pulp and saccharification of recycled paper sludge. *Bioresour. Technol.* **125**, 182–187 (2012).
7. D. Gao, *et al.*, Hemicellulases and auxiliary enzymes for improved conversion of lignocellulosic biomass to monosaccharides. *Biotechnol. Biofuels* **4** (2011).
8. R. Goldbeck, *et al.*, Development of hemicellulolytic enzyme mixtures for plant biomass deconstruction on target biotechnological applications. *Appl. Microbiol. Biotechnol.* **98**, 8513–8525 (2014).
9. D. Zarafeta, *et al.*, Discovery and Characterization of a Thermostable and Highly Halotolerant GH5 Cellulase from an Icelandic Hot Spring Isolate. *PLoS One* **11**, e0146454 (2016).
10. K. S. Chuan Wei, T. C. Teoh, P. Koshy, I. Salmah, A. Zainudin, Cloning, expression and characterization of the endoglucanase gene from Bacillus subtilis UMC7 isolated from the gut of the indigenous termite Macrotermes malaccensis in Escherichia coli. *Electron. J. Biotechnol.* **18**, 103–109 (2015).
11. S. Vidal-Melgosa, *et al.*, A new versatile microarray-based method for high-throughput screening of carbohydrate-active enzymes. *J. Biol. Chem.* **290**, 9020–9036 (2015).
12. M. B. Jaffee, B. Imperiali, Optimized protocol for expression and purification of membrane-bound PglB, a bacterial oligosaccharyl transferase. *Protein Expr. Purif.* **89**, 241–250 (2013).
13. J. Tiralongo, A. Maggioni, "The Targeted Expression of Nucelotide Sugar Transporters to the E.coli Inner Membrane" in *Heterologous Gene Expression in E.Coli*, (2010), pp. 237–249.
14. N. Nelson, A photometric adaptation of the SOMOGYI method for the determination of glucose. *J. Biol. Chem.* **03**, 375–380 (1944).

Annex 2: Supplementary material of Chapter 2

Text S1: Genes selection, isolation and cloning

Following the bioinformatics analyses, genes of interest were identified and amplified by polymerase chain reaction (PCR), using a Veriti™ 96 wells Thermal cycler (Applied Biosystems, Foster City, USA). Amplicons were generated using the Q5 High-Fidelity Hot Start (New England Biolabs Inc., Ipswich, USA) and according to the thermal profile: initial denaturation was made at 98°C for 5 seconds, followed by annealing for 30 seconds with temperature depending on the melting temperature of the primers (Tm), and the final extension was carried out at 72°C for 30 seconds. Thirty six cycles of PCR were performed. PCR products were purified using the PCR purification kit (Qiagen, Hilden, Germany) and concentrations were determined by NanoDrop 1000 Spectrometer (Thermofisher, Waltham, USA).

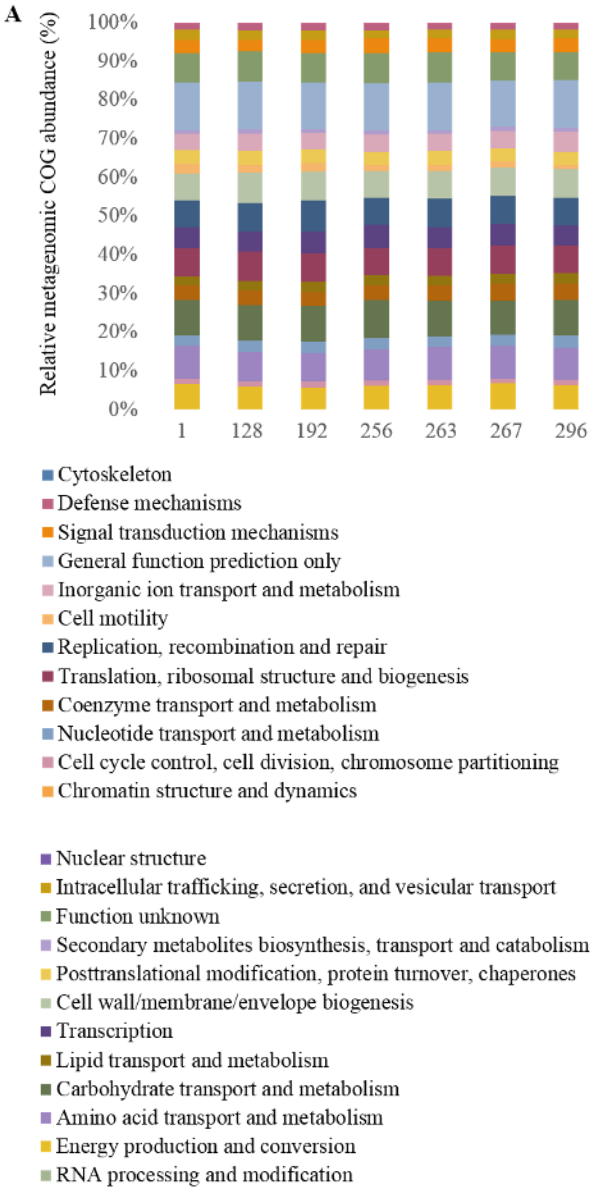
Genes were then cloned into pGem-t-easy vector using the respective kit (Promega, Madison, USA) followed by transformation of JM109 High Efficiency Competent Cells, allowing blue/white screening of colonies. Positive colonies were confirmed by a colony PCR and the respective plasmids from overnight cultures were purified using GeneJET Plasmid Miniprep Kit (Thermofisher, Waltham, USA). Correctly inserted genes (determined by sequencing) were selected for further expression.

Text S2: Proteins identification

Recombinant proteins were identified after SDS-PAGE separation. The band at the expected molecular weight was cut from the gel and reduced with 100 µL of 0.15% (w/v) D-Dithiothreitol (DDT) in 100 mM Ammonium bicarbonate (AmBic). After incubation at 56°C for 30 minutes, DDT was removed and 100 µL of 1% (w/v) Iodoacetamide in 100 mM AmBic was added to the sample. Incubation was performed for 30 min at room temperature in the dark. Iodoacetamide was removed from the sample with two washes with 200 µL 50 mM AmBic in ethanol. The gel pieces were dehydrated with 200 µL of acetonitrile (ACN) and after removal of the liquid phase, the gel pieces were completely dried. The dried gel pieces were incubated at 37°C overnight in 8 µL of trypsin (5ng/µL) and 15 µL 50mM AmBic. The next day, the liquid phase containing the peptides was recovered in a clean vial and peptides were further extracted by two washes with 35 µL 50% (v/v) ACN, 0.1% (v/v) Trifluoroacetic acid (TFA). The extracted peptides were dried under vacuum for 2 hours. Peptides were re-solubilized in 0.7 µL 50% (v/v) ACN, 0.1% (v/v) TFA and spotted on a stainless steel OptiTOF 384 well plate (AB Sciex, Framingham, USA). 0.7 µL of α-Cyano-4-hydroxycinnamic acid (CHCA) matrix at 7mg/mL in 50% (v/v) ACN, 0.1% (v/v) TFA were added. Samples were analysed using the AB Sciex 5800 MALDI-TOF/TOF (Sciex, Framingham, USA), a MS spectrum was acquired and the ten most intense peaks, excluding known contaminants, were fragmented. Identification was performed using an in-house MASCOT server with a database containing the theoretical amino acid sequence of the recombinant proteins and standard settings (B. Printz, K. Sergeant, S. Lutts, C. Guignard, J. Renaut, and JF. Hausman, J Proteome Res 12:5160–5179, 2013, doi:10.1021/pr400590d). Additional *de novo* sequencing was done to confirm the presence of any signal peptides.

Table S1: Concentrations of the substrates used for the enzymatic assays

Substrate	Stock concentration (mM)
4-nitrophenyl α-D-mannopyranoside	20
4-nitrophenyl β-D-glucopyranoside	40
4-nitrophenyl β-D-xylopyranoside	40
4-nitrophenyl β-D-mannopyranoside	20
4-nitrophenyl α-L-arabinofuranoside	20
4-nitrophenyl β-D-cellobioside	10
4-nitrophenyl α-D-galactopyranoside	40
4-nitrophenyl acetate	4
4-nitrophenyl α-D-glucopyranoside	20
Substrate	**Stock concentration (g/L)**
CMC	10
Arabinoxylan	12
Galactomannan	8
Glucomannan	8
Xylan	18


A
Relative metagenomic COG abundance (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
128
192
256
263
267
296
Cytoskeleton
Defense mechanisms
Signal transduction mechanisms
General function prediction only
Inorganic ion transport and metabolism
Cell motility
Replication, recombination and repair
Translation, ribosomal structure and biogenesis
Coenzyme transport and metabolism
Nucleotide transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Chromatin structure and dynamics
Nuclear structure
Intracellular trafficking, secretion, and vesicular transport
Function unknown
Secondary metabolites biosynthesis, transport and catabolism
Posttranslational modification, protein turnover, chaperones
Cell wall/membrane/envelope biogenesis
Transcription
Lipid transport and metabolism
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Energy production and conversion
RNA processing and modification


B
Relative metagenomic abundance of genes assigned to KOs involved in acetogenesis (%)
K15023
K15022
K14138
K05299
K01938
K01491
K00297
K00198
K00197
K00194
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
128
192
256
263
267
296
C
Relative metagenomic abundance of genes assigned to KOs involved in methanogenesis (%)
K00925
K01895
K00200
K00202
K00320
K08264
K00580
K03390
K00584
K00125
K00577
K00205
K00625
K03389
K00203
K14127
K11261
K00401
K00579
K13788
K00197
K00581
K00193
K08265
K03388
K00201
K00672
K01499
K14126
K00399
K00578
K00319
K00194
K14128
K00402
D
Relative metagenomic CAZy abundance (%)
PL
GT
GH
CE
CBM
AA
Time (days)
E
Relative metagenomic CBM abundance (%)
CBM9
CBM70
CBM61
CBM54
CBM50
CBM41
CBM37
CBM30
CBM22
CBM17
CBM78
CBM67
CBM6
CBM53
CBM48
CBM40
CBM35
CBM28
CBM21
CBM16
CBM77
CBM66
CBM57
CBM52
CBM47
CBM4
CBM34
CBM25
CBM20
CBM13
CBM76
CBM62
CBM56
CBM51
CBM44
CBM38
CBM32
CBM23
CBM2
CBM12
Time (days)

Figure S1: Metagenomics-assisted characterisation of the functional community potential during the anaerobic digestion (AD) experiment. **A** - Relative metagenomic abundance of the different clusters of orthologous groups (COG) coloured according to the COG group. **B** - Relative metagenomic abundance of the different Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) involved in the acetogenesis pathway coloured according to the KO number. **C** - Relative metagenomics abundance of the different KEGGs involved in the methanogenesis (including acetoclastic and hydrogenotrophic pathways) coloured according to the KO number. **D** - Relative metagenomic abundance of the different CAZymes families coloured according to the CAZyme family. The differents CAZymes families represented are; auxiliary enzyme (AA); Carbohydrate binding module (CBM); Carbohydrate esterase (CE); Glycoside hydrolase (GH); Glycosyltransferase (GT); Polysaccharide lyase (PL). **E** - Relative metagenomic abundance of the different CBM families coloured according to the CBM family. Note: Small variability of the functional profile of genes involved in methanogenesis was related to the overall decrease of genes encoding for the heterodisulfide reductase (hydrogenotrophic pathway) during acidosis.

Table S2: Analysis of the different MAGs. The percentage of CAZys in the MAGs refers to the proportion of putative CAZymes from all genes identified within the MAG

Phylophlan	MAG number	Proteins coding genes	sus-like genes	CAZys coding genes	% CAZys in the MAG	% Completeness	% Metagenomic abundance
Bacteria; Bacteroidetes	2	2263	97	222	9.8	96.6	6.3
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales	8	1515	41	151	10.0	82.9	8.8
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	11	1881	29	151	8.0	85.9	1.2
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales	15	1461	23	135	9.1	81.6	12
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	17	1320	25	62	4.7	91.2	1.2
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	25	1276	32	104	8.2	59	0.5
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales	26	1237	34	107	8.7	51.3	0.3
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales	28	1726	60	160	9.3	54.5	0.6
Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	30	1914	9	228	11.9	99.5	3.4
Bacteria; Firmicutes; Clostridia; Clostridiales	3	1784	0	73	4.1	95.3	1.6
Bacteria; Firmicutes; Clostridia	4	2099	1	89	4.2	94.9	2.6
Bacteria; Firmicutes; Clostridia	12	2064	0	74	3.6	89.8	2.4
Bacteria; Firmicutes; Clostridia	16	1615	0	74	4.6	63.9	0.7
Bacteria; Firmicutes; Clostridia	20	1307	1	45	3.4	60.4	0.5
Bacteria; Firmicutes; Clostridia; Clostridiales	22	1040	0	58	5.6	62.0	0.3
Bacteria; Firmicutes; Clostridia; Clostridiales	23	1034	0	43	4.2	51.8	0.6
Bacteria; Firmicutes; Clostridia	27	1216	0	27	2.2	49.5	0.3
Bacteria; Cloacimonetes	1	1250	3	40	3.2	98.9	22.3
Bacteria; Cloacimonetes	13	1404	2	62	4.4	87.3	1.1
Bacteria; Cloacimonetes	31	1443	2	60	4.2	95.6	1.7
Bacteria; Spirochaetes; Spirochaetales; Spirochaetales	7	1945	1	90	4.6	66.7	1.1
Bacteria; Spirochaetes; Spirochaetales; Spirochaetales	14	1889	0	62	3.3	83.8	0.8
Bacteria; Spirochaetes; Spirochaetales; Spirochaetales; Treponemaceae; Treponema	18	1621	0	47	2.9	76.3	0.6
Bacteria; Synergistetes; Synergistia; Synergistales; Synergistaceae	5	1790	1	68	3.8	94.8	1.5
Bacteria; Synergistetes; Synergistia; Synergistales; Synergistaceae	9	1708	2	44	2.6	92.8	0.7
Bacteria; Synergistetes; Synergistia; Synergistales; Synergistaceae	10	1368	2	29	2.1	92.2	1
Bacteria; Lentisphaerae; Lentisphaerae	19	1736	0	113	6.5	77.3	1.4
Bacteria; Planctomycetes; Planctomycetacia	21	1935	6	171	8.8	65.5	0.6
Bacteria; Thermotogae; Thermotogae; Thermotogales; Thermotogaceae	24	1273	0	78	6.1	66.8	0.4
Archaea; Euryarchaeota; Methanomicrobia; Methanomicrobiales	6	1447	0	28	1.9	92.2	1.5

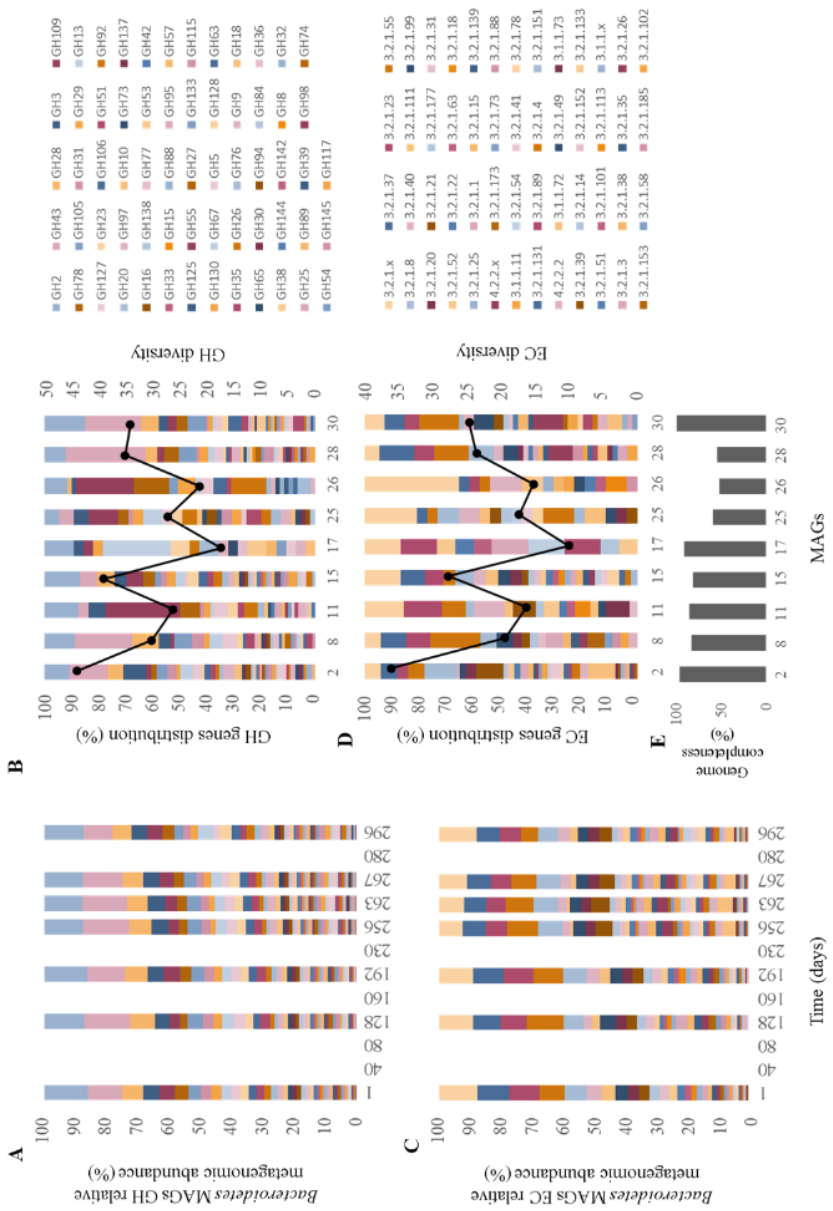

A
Bacteroidetes MAGs GH relative metagenomic abundance (%)
B
GH diversity
GH genes distribution (%)
C
Bacteroidetes MAGs EC relative metagenomic abundance (%)
D
EC diversity
EC genes distribution (%)
E
Genome completeness (%)
MAGs
Time (days)

Figure S2: Metagenomics-assisted characterisation of *Bacteroidetes* metagenome assembles genomes (MAGs) and their carbohydrate hydrolytic potential throughout the anaerobic digestion (AD) experiment. **A** - Relative metagenomic abundance of the different glycoside hydrolases (GHs) coloured according to the assigned GH family and throughout the digestion experiment. **B** - Distribution of the annotated GH genes for the different *Bacteroidetes* MAGs, coloured according to the assigned GH family. The number of assigned GH families (i.e. GH diversity) is represented by a black solid line. **C** - Relative metagenomic abundance of the different GHs additionally functionally assigned to an EC category and coloured according to the EC category. **D** - Distribution of the assigned EC categories among the different *Bacteroidetes* MAGs, coloured according to the EC category and throughout the digestion experiment. The number of EC functional categories (i.e. EC diversity) is represented by a black solid line. **E** - Percentage of genome completeness for the different *Bacteroidetes* MAGs.
Remark: Due to incompleteness of some *Bacteroidetes* MAGs, the diversity measurements for the different GH families (respectively EC categories) can be higher than shown on the respective graphs.

Table S3: Results of the enzymatic assays and EC number prediction

		Activity (10^{-3} µmol/min /mg ± SD)					
		uBac-GH26a	uBac-GH26b[+]	uBac-GH130	uBac-GH5[++]	uBac-CE6[-]	uBac-GH3
Substrate type	Substrate	EC.3.2.1.78*	EC.3.2.1.78*	EC. 2.4.1.281*	EC. 3.2.1.4*	NA	EC.3.2.1.37*
Polysaccharide	Xylan	-	-	-	-	-	-
	Galactomannan	94.86 ± 6.08	90.68 ± 5.28	-	-	-	-
	Arabinoxylan	-	-	-	-	-	-
	Carboxymethyl cellulose	-	-	-	19.05 ± 0.51	-	-
	Glucomannan	102.84 ± 1.06	31.39 ± 1.52	-	-	-	-
4-Nitrophenyl	α-D-mannopyranoside	-	-	-	-	-	-
	β-D-glucopyranoside	-	-	-	-	-	177.00 ± 7.52
	β-D-xylopyranoside	-	-	-	-	-	-
	β-D-mannopyranoside	-	-	-	-	-	-
	α-L-arabinofuranoside	-	-	-	-	-	-
	β-D-cellobioside	-	-	-	9.94 ± 0.88	-	-
	α-D-galactopyranoside	-	-	-	-	-	-
	acetate	-	-	-	-	30215.70 ± 0.20	-
	α-D-glucopyranoside	-	-	-	-	-	-

Results are expressed in specific activity (µmol/min/mg), i.e. the amount of substrate converted per minute and milligram of protein.
* EC number prediction based on Hotpep analysis
- No activity detected
[N] Not assigned
[+] Signal peptidase I and cleavage site predicted using LipoP v1.0 server
[++] Lipoprotein signal peptide II and cleavage site predicted using LipoP v1.0 server

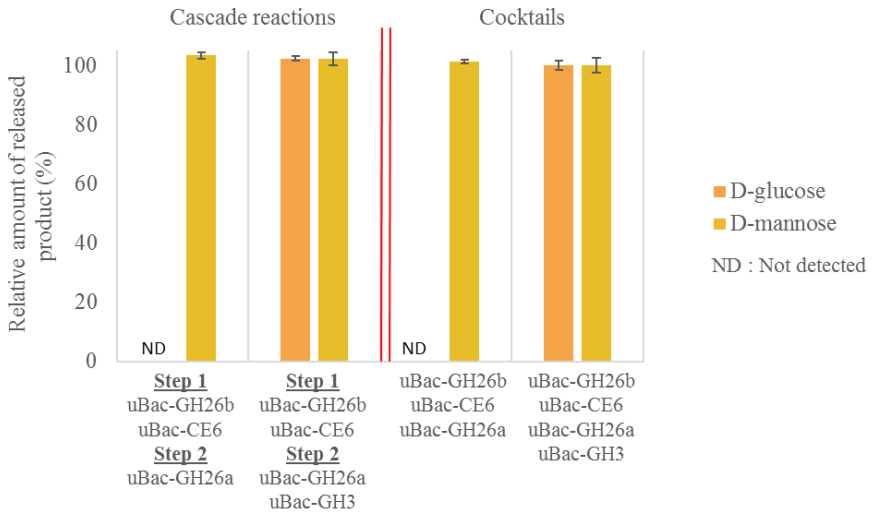
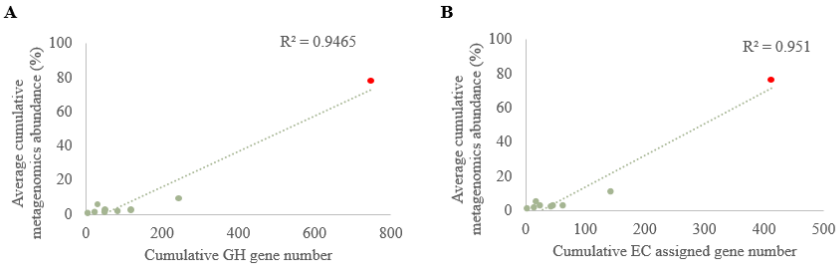


Figure S3: Release of D-glucose and D-mannose after enzymatic hydrolysis of acetylated konjac glucomannan. Cascade reactions and cocktails were compared. Cascade reaction is defined as follow, first hydrolysis of the substrate (one hour at 37°C) by the enzymes specified in step 1, followed by a second hydrolysis (one hour at 37°C) by the enzyme(s) specified in step 2. Cocktail is defined as follow, enzymes were added and a single incubation step (one hour at 37°C) was performed in order to hydrolyse the substrate.

Figure S4: Correlation between the average cumulative metagenomic abundance of metagenome assembled genomes (MAGs) represented at the phylum level and the



cumulative number of their carbohydrate hydrolytic genes, shown at the glycoside hydrolase (GH) family level **(A),** and further supported by the assigned EC category **(B)**. *Bacteroidetes* phylum is represented by a red dot.

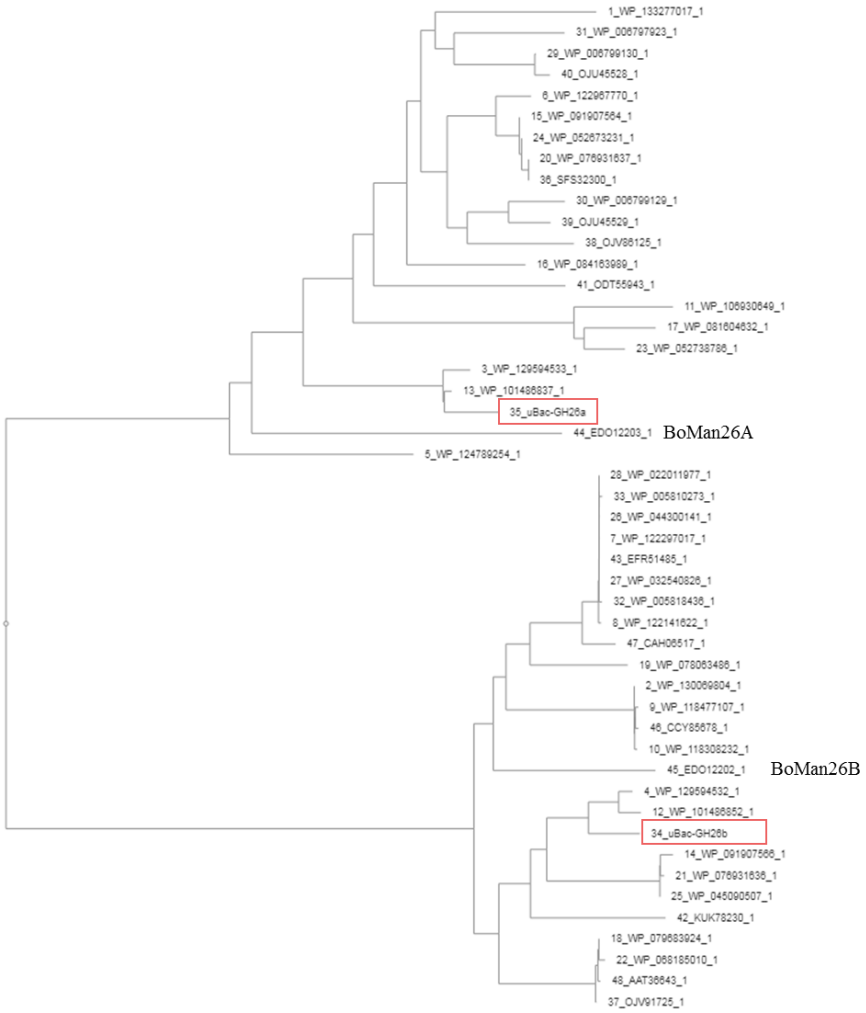
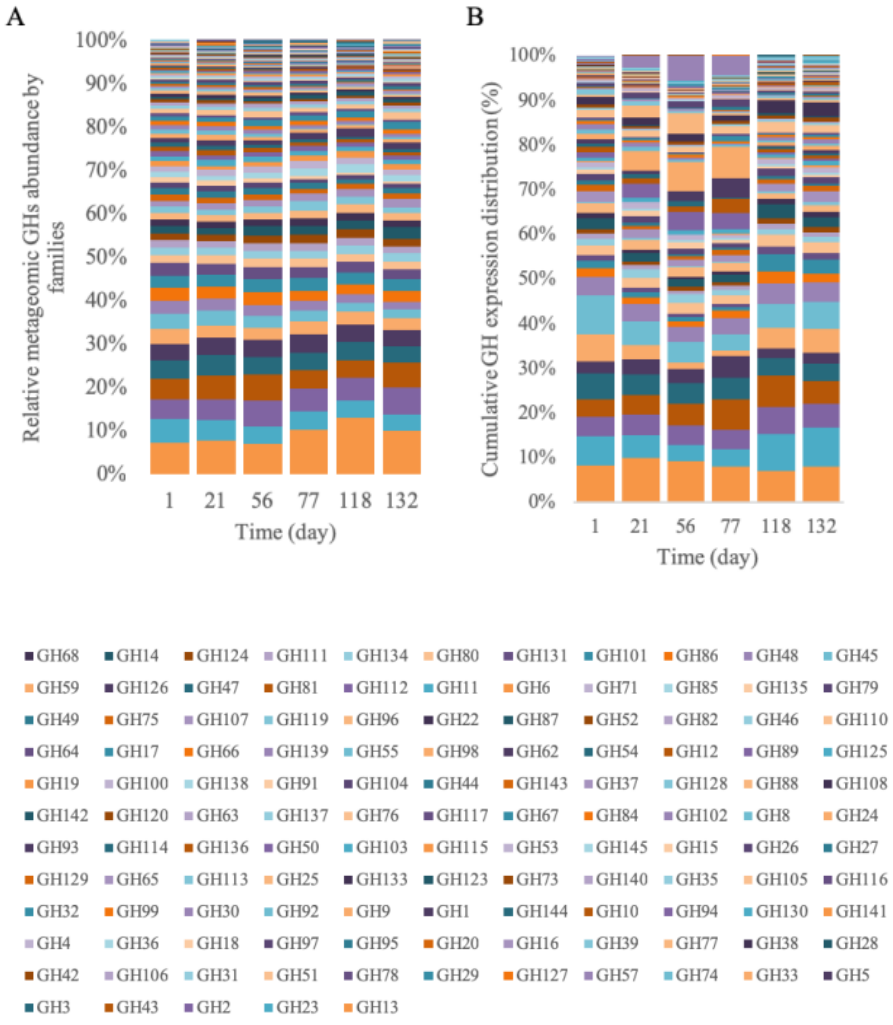


Figure S5: uBac-GH26a and uBac-GH26b sequence similarities including mannanases from NCBI database. Sequence alignment and phylogenetic tree was generated using the MAFFT server https://mafft.cbrc.jp/alignment/software. BoMan26A and BoMan26B refers to two mannanases isolated from a galactomannan-targeting PUL (Bagenholm et al., 2017). uBac-GH26a and uBac-GH26b are highlighted in red.

Bagenholm, V., Reddy, S. K., Bouraoui, H., Morrill, J., Kulcinskaja, E., Bahr, C. M., et al. (2017). Galactomannan catabolism conferred by a polysaccharide utilization locus of Bacteroides ovatus: Enzyme synergy and crystal structure of a β-mannanase. Journal of Biological Chemistry, 292(1), 229–243. https://doi.org/10.1074/jbc.M116.746438

Annex 3: Supplementary material of Chapter 3



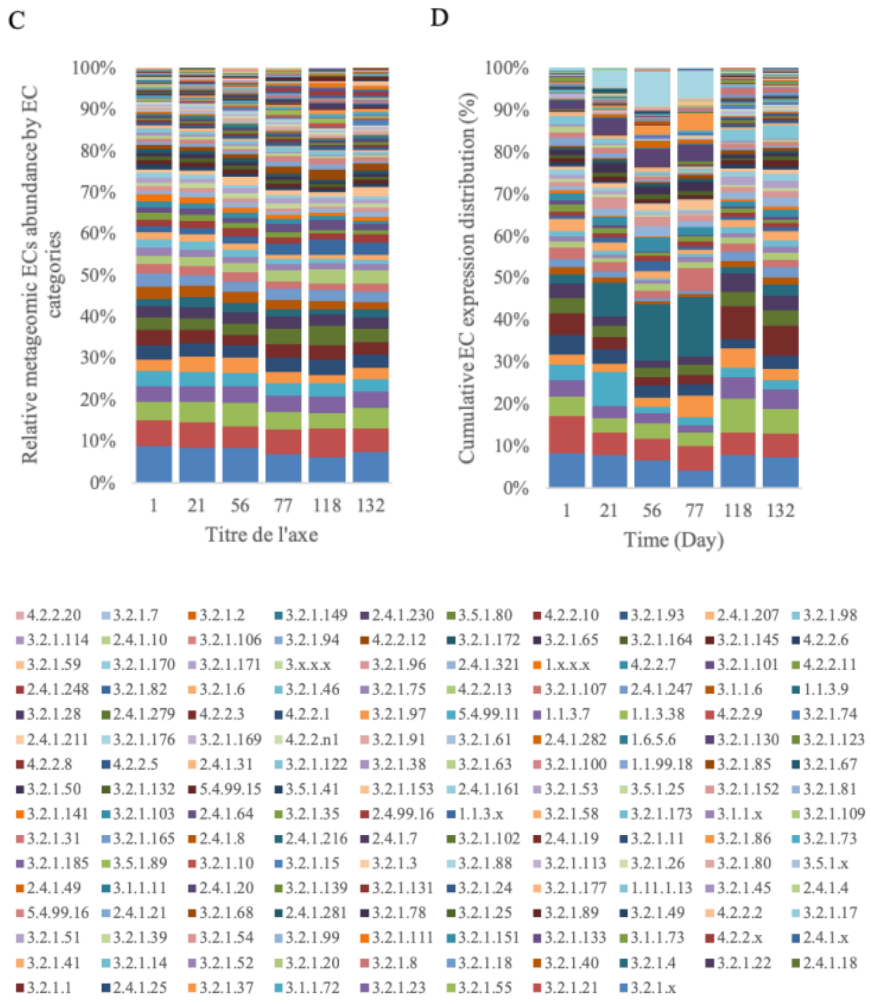
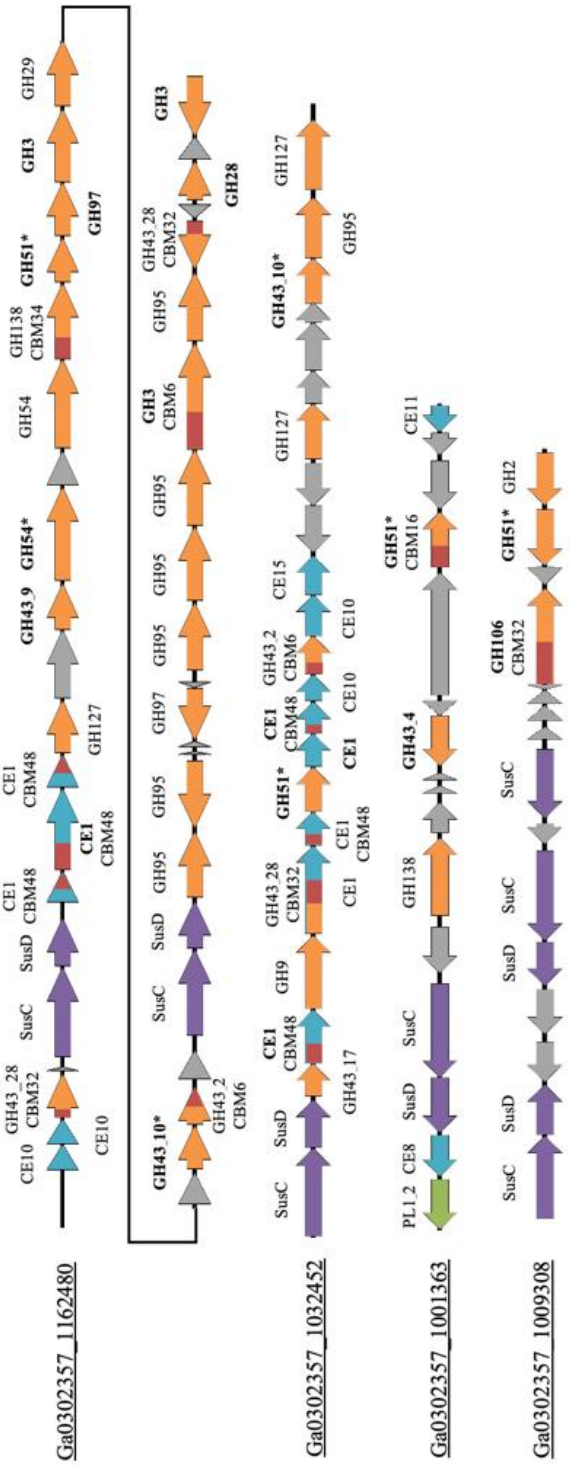


Figure S1: Glycoside Hydrolases (GHs) functional redundancy and plasticity from anaerobic digestion (AD) microbiome fed with sugar beet pulp. **A** - Relative metagenomic abundance of the different GH families at the microbiome level. **B** - Cumulative expression of GH coding genes at the microbiome level. **C** - Relative metagenomics abundance of the different GHs further assigned to an Enzyme Commission (EC) category, at the microbiome level. **D** - Cumulative expression of GH coding genes further assigned to an EC category, at the microbiome level.

Table S1: Results of the calculated genome average nucleotide idendity (ANI) as an indicator of the metagenome assembled genomes (MAGs) novelty

Phylum	**Number of MAGs displaying …**		
	ANI > 95%	95%>ANI>75%	75% > ANI
Acidobacteria	3	4	3
Actinobacteria	8	8	2
Bacteroidetes	16	4	14
Chloroflexi	7	13	14
Firmicutes	7	7	20
Planctomycetes	0	5	0
Proteobacteria	6	14	21
Spirochaetes	1	2	3
Synergistetes	3	2	2
Verrucomicrobia	2	3	4
Total	53	62	83


Ga0302357_1162480
CE10
GH43_28 CBM32
SusC
SusD
CE1 CBM48
CE1 CBM48
CE1 CBM48
GH127
GH43_9
GH54*
GH54
GH138 CBM34
GH51*
GH3
GH29
GH97
GH43_10*
GH43_2 CBM6
SusC
SusD
GH95
GH95
GH97
GH95
GH95
GH95
GH3 CBM6
GH43_28 CBM32
GH28
GH3
Ga0302357_1032452
SusC
SusD
CE1 CBM48
GH43_17
GH9
GH43_28 CBM32
GH51*
CE1 CBM48
CE1
CE1 CBM48
CE1
CE10
GH43_2 CBM6
CE15
GH127
GH43_10*
GH95
GH127
Ga0302357_1001363
PL1_2
CE8
SusD
SusC
GH138
GH43_4
GH51* CBM16
CE11
Ga0302357_1009308
SusC
SusD
SusD
SusC
SusC
GH106 CBM32
GH51*
GH2

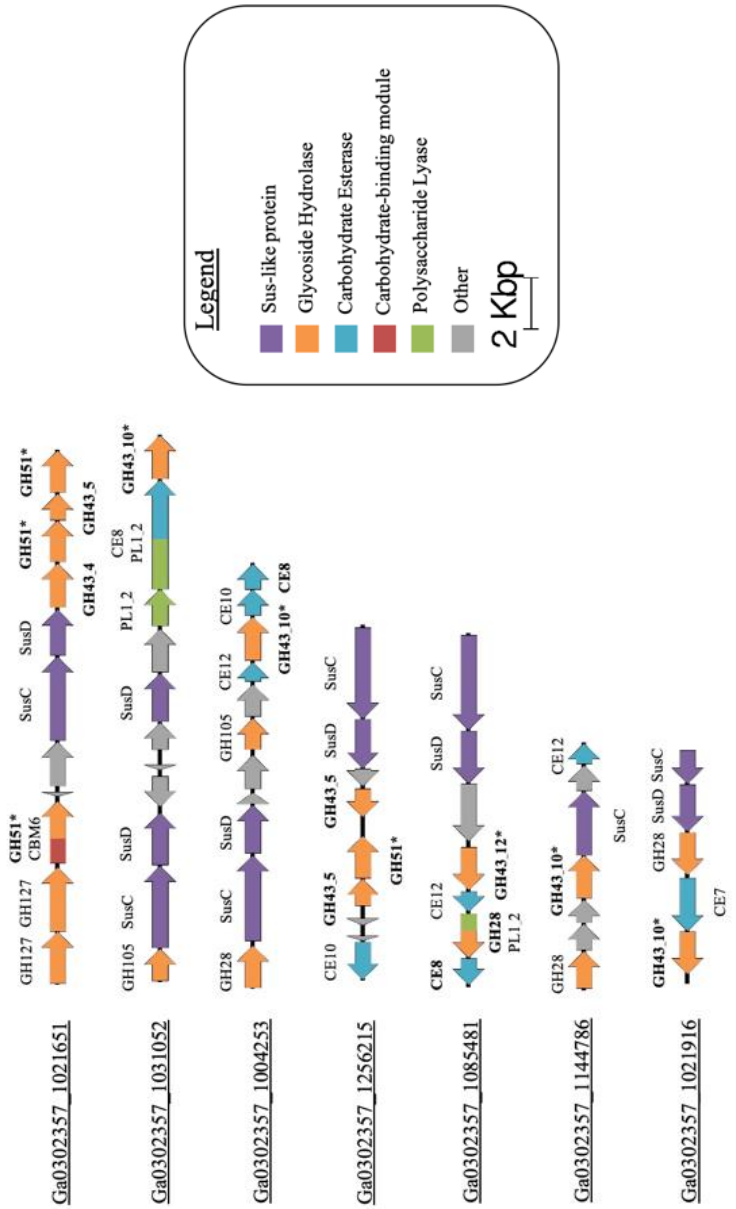


Figure S2: Expressed polysaccharide utilization loci (PULs) harbouring a-L-arabinofuranosidase (EC. 3.2.1.55) activity and identified from the anaerobic digestion microbiome fed with sugar beet pulp. Further assignement to an EC category is represented by a bold font, and asterix identifies the putative a-L-arabinofuranosidase coding gene (further assigned to EC 3.2.1.55).

Table S2: Identified putative polysaccharide utilization loci targeting (glucurono)arabinoxylan from anaerobic digester microbiome

PUL	Gene ID	CAZy family // Sus-like gene	EC number
Ga0302357_1045369	Ga0302357_10453693	GH30	
	Ga0302357_10453694	GH130	
	Ga0302357_10453695	Other	
	Ga0302357_10453696	Other	
	Ga0302357_10453697	CE6	
	Ga0302357_10453698	Other	
	Ga0302357_10453699	CE1	
	Ga0302357_104536910	Other	
	Ga0302357_104536911	SusD	
	Ga0302357_104536912	SusC	
	Ga0302357_104536913	Other	
	Ga0302357_104536914	Other	
	Ga0302357_104536915	Other	
	Ga0302357_104536916	CBM6_GH43_2	
	Ga0302357_104536917	CBM30_GH9	
	Ga0302357_104536918	GH43_10	3.2.1.8
	Ga0302357_104536919	GH95	
	Ga0302357_104536920	GH67	3.2.1.139
	Ga0302357_104536921	CE6	
	Ga0302357_104536922	GH51	
	Ga0302357_104536923	GH43_10	3.2.1.55
	Ga0302357_104536924	GH43_35	
	Ga0302357_104536925	CBM51_GH27	3.2.1.88
Ga0302357_1004978	Ga0302357_100497817	GH3	3.2.1.37
	Ga0302357_100497818	CBM48_CE1	3.1.1.72
	Ga0302357_100497819	Other	
	Ga0302357_100497820	GH43	
	Ga0302357_100497821	GH115	3.2.1.131
	Ga0302357_100497822	GH43_2_CBM6	
	Ga0302357_100497823	GH43_10	3.2.1.55
	Ga0302357_100497824	GH67	3.2.1.139
	Ga0302357_100497825	CBM6_GH43_29	3.2.1.37
	Ga0302357_100497826	GH5_21	3.2.1.8
	Ga0302357_100497827	GH27_CBM51	
	Ga0302357_100497828	GH43_12	3.2.1.37

	Ga0302357_100497829	CE1	3.1.1.73
	Ga0302357_100497830	CE6	3.1.1.72
	Ga0302357_100497831	GH43_1	3.2.1.37
	Ga0302357_100497832	GH10	3.2.1.8
	Ga0302357_100497833	Other	
	Ga0302357_100497834	SusD	
	Ga0302357_100497835	SusC	
Ga0302357_1111851	Ga0302357_11118512	GH27_CBM51	
	Ga0302357_11118513	GH5_21	3.2.1.8
	Ga0302357_11118514	CBM6_GH43_29	3.2.1.37
	Ga0302357_11118515	GH43_29_CBM6	3.2.1.37
	Ga0302357_11118516	CE6	3.1.1.72
	Ga0302357_11118517	CE1_CBM6	
	Ga0302357_11118518	GH127	
	Ga0302357_11118519	GH31	
	Ga0302357_111185110	GH43_10	3.2.1.55
	Ga0302357_111185111	Other	
	Ga0302357_111185112	SusC	
	Ga0302357_111185113	SusD	
	Ga0302357_111185114	SusC	
	Ga0302357_111185115	SusD	
Ga0302357_1046105	Ga0302357_10461051	CBM6	
	Ga0302357_10461052	GH67	3.2.1.139
	Ga0302357_10461053	Other	
	Ga0302357_10461054	Other	
	Ga0302357_10461055	SusC	
	Ga0302357_10461056	SusD	
	Ga0302357_10461057	GH115	3.2.1.131
	Ga0302357_10461058	GH115	3.2.1.131
	Ga0302357_10461059	GH43	
	Ga0302357_104610510	CBM51	
	Ga0302357_104610511	Other	
	Ga0302357_104610512	GH51_CBM22	3.2.1.55
Ga0302357_1054068	Ga0302357_105406845	GH43_1	3.2.1.37
	Ga0302357_105406846	GH10	3.2.1.8
	Ga0302357_105406847	Other	
	Ga0302357_105406848	SusC	

	Ga0302357_105406849	Other	
	Ga0302357_105406850	Other	
	Ga0302357_105406851	Other	
	Ga0302357_105406852	CE8_PL1_2	
	Ga0302357_105406853	Other	
	Ga0302357_105406854	CE10	
	Ga0302357_105406855	Other	
	Ga0302357_105406856	Other	
	Ga0302357_105406857	Other	
	Ga0302357_105406858	Other	
	Ga0302357_105406859	GH2	
	Ga0302357_105406860	Other	
	Ga0302357_105406861	GH51_CBM4	3.2.1.55
Ga0302357_1081478	Ga0302357_10814785	GH31	
	Ga0302357_10814786	CE1_CBM48	3.1.1.72
	Ga0302357_10814787	CE1_CBM48	3.1.1.72
	Ga0302357_10814788	Other	
	Ga0302357_10814789	CBM51	
	Ga0302357_108147810	GH43_29_CBM6	3.2.1.37
	Ga0302357_108147811	Other	
	Ga0302357_108147812	Other	
	Ga0302357_108147813	Other	
	Ga0302357_108147814	Other	
	Ga0302357_108147815	GH43_2_CBM6	
	Ga0302357_108147816	GH43_10	3.2.1.55
	Ga0302357_108147817	GH10	3.2.1.8
	Ga0302357_108147818	CBM48_CE1	3.1.1.73
	Ga0302357_108147819	SusD	
Ga0302357_1010789	Ga0302357_10107892	CBM6_GH43_2	
	Ga0302357_10107893	GH43	3.2.1.55
	Ga0302357_10107894	CBM48_CE1	3.1.1.72
	Ga0302357_10107895	SusD	
	Ga0302357_10107896	SusD	
	Ga0302357_10107897	SusC	
	Ga0302357_10107898	CE6_GH43_29	3.2.1.8
	Ga0302357_10107899	CE1	3.1.1.73
Ga0302357_1005151	Ga0302357_10051514	CE10	

	Ga0302357_10051515	SusC	
	Ga0302357_10051516	SusD	
	Ga0302357_10051517	CBM48_CE1	
	Ga0302357_10051518	GH3	3.2.1.21
	Ga0302357_10051519	GH51	3.2.1.55

Table S3: Identified putative polysaccharide utilization loci targeting xyloglucan and pectin from anaerobic digester microbiome

Contig	Gene ID	CAZY_family	EC number
Ga0302357_1001363	Ga0302357_100136313	PL1_2	
	Ga0302357_100136314	CE8	
	Ga0302357_100136315	SusD	
	Ga0302357_100136316	SusC	
	Ga0302357_100136317	Other	
	Ga0302357_100136318	GH138	
	Ga0302357_100136319	Other	
	Ga0302357_100136321	Other	
	Ga0302357_100136322	Other	
	Ga0302357_100136323	GH43_4	3.2.1.99
	Ga0302357_100136324	Other	
	Ga0302357_100136325	Other	
	Ga0302357_100136326	GH51_CBM16	3.2.1.55
	Ga0302357_100136327	Other	
	Ga0302357_100136328	Other	
	Ga0302357_100136329	CE11	
Ga0302357_1009308	Ga0302357_10093081	SusC	
	Ga0302357_10093082	SusD	
	Ga0302357_10093083	Other	
	Ga0302357_10093084	Other	
	Ga0302357_10093085	SusD	
	Ga0302357_10093086	SusC	
	Ga0302357_10093087	Other	
	Ga0302357_10093088	SusC	
	Ga0302357_10093089	Other	
	Ga0302357_100930810	Other	
	Ga0302357_100930811	Other	
	Ga0302357_100930812	Other	

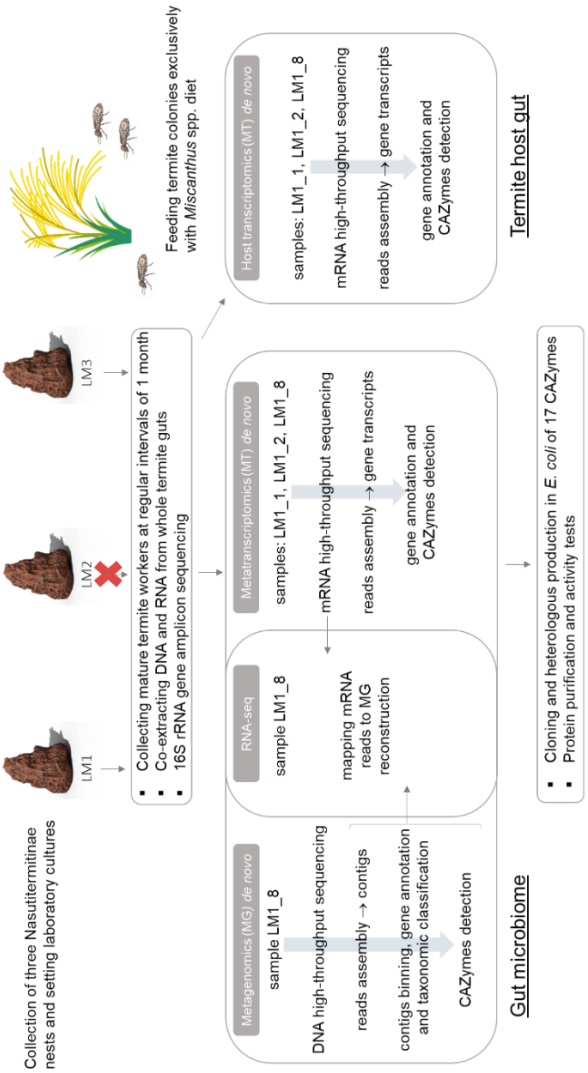
	Ga0302357_100930813	CBM32_GH106	3.2.1.40
	Ga0302357_100930814	Other	
	Ga0302357_100930815	GH51	3.2.1.55
	Ga0302357_100930816	GH2	
Ga0302357_1021651	Ga0302357_102165119	GH127	
	Ga0302357_102165120	GH127	
	Ga0302357_102165121	CBM16_GH51	3.2.1.55
	Ga0302357_102165122	Other	
	Ga0302357_102165123	Other	
	Ga0302357_102165124	SusC	
	Ga0302357_102165125	SusD	
	Ga0302357_102165126	GH43_4	3.2.1.99
	Ga0302357_102165127	GH51	3.2.1.55
	Ga0302357_102165128	GH43_5	3.2.1.99
	Ga0302357_102165129	GH51	3.2.1.55
Ga0302357_1031052	Ga0302357_10310524	GH105	
	Ga0302357_10310525	SusC	
	Ga0302357_10310526	SusD	
	Ga0302357_10310527	Other	
	Ga0302357_10310528	Other	
	Ga0302357_10310529	Other	
	Ga0302357_103105210	SusD	
	Ga0302357_103105211	Other	
	Ga0302357_103105212	PL1_2	
	Ga0302357_103105213	CE8_PL1_2	
	Ga0302357_103105214	GH43_10	3.2.1.55
Ga0302357_1004253	Ga0302357_10042535	GH28	
	Ga0302357_10042536	SusC	
	Ga0302357_10042537	SusD	
	Ga0302357_10042538	Other	
	Ga0302357_10042539	Other	
	Ga0302357_100425310	GH105	
	Ga0302357_100425311	Other	
	Ga0302357_100425312	CE12	
	Ga0302357_100425313	GH43_10	3.2.1.55
	Ga0302357_100425314	CE10	
	Ga0302357_100425315	CE8	3.1.1.11

Ga0302357_1256215	Ga0302357_12562158	CE10	
	Ga0302357_12562159	Other	
	Ga0302357_125621510	Other	
	Ga0302357_125621511	GH43_5	3.2.1.99
	Ga0302357_125621512	GH51	3.2.1.55
	Ga0302357_125621513	GH43_5	3.2.1.99
	Ga0302357_125621514	Other	
	Ga0302357_125621515	SusD	
	Ga0302357_125621516	SusC	
Ga0302357_1085481	Ga0302357_10854812	CE8	3.1.1.11
	Ga0302357_10854813	PL1_2_GH28	3.2.1.15
	Ga0302357_10854814	CE12	
	Ga0302357_10854815	GH43_10	3.2.1.55
	Ga0302357_10854816	Other	
	Ga0302357_10854817	SusD	
	Ga0302357_10854818	SusC	
Ga0302357_1144786	Ga0302357_11447864	GH28	
	Ga0302357_11447865	Other	
	Ga0302357_11447866	Other	
	Ga0302357_11447867	GH43_10	3.2.1.55
	Ga0302357_11447868	SusC	
	Ga0302357_11447869	Other	
	Ga0302357_114478610	CE12	
Ga0302357_1021916	Ga0302357_10219162	GH43_10	3.2.1.55
	Ga0302357_10219163	CE7	
	Ga0302357_10219164	GH28	
	Ga0302357_10219165	SusD	
Ga0302357_1032452	Ga0302357_10324523	SusC	
	Ga0302357_10324524	SusD	
	Ga0302357_10324525	GH43_17	
	Ga0302357_10324526	CBM48_CE1	3.1.1.72
	Ga0302357_10324527	GH9	
	Ga0302357_10324528	GH43_28_CBM32	
	Ga0302357_10324529	CBM48_CE1	
	Ga0302357_103245210	GH51	3.2.1.55
	Ga0302357_103245211	CE1	3.1.1.72
	Ga0302357_103245212	CE1_CBM48	3.1.1.72

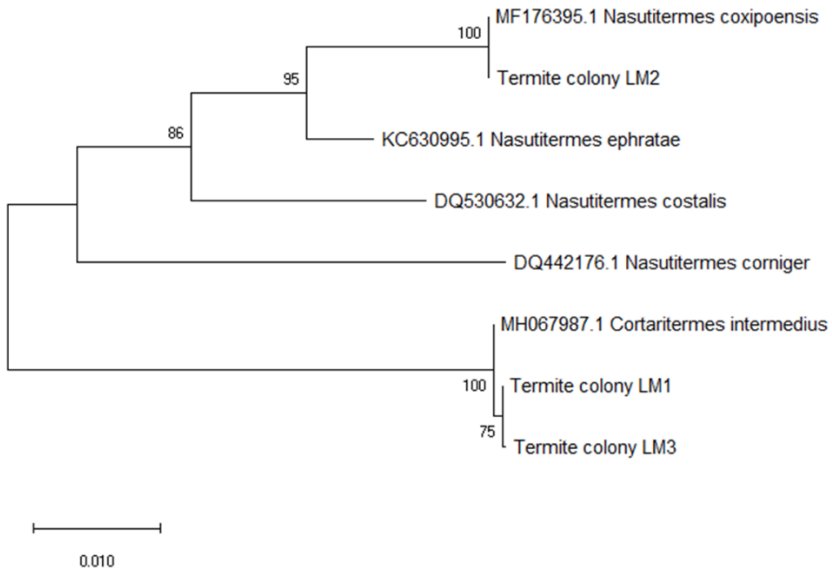
	Ga0302357_103245213	CE10	
	Ga0302357_103245214	CBM6_GH43_2	
	Ga0302357_103245215	CE10	
	Ga0302357_103245216	CE15	
	Ga0302357_103245217	Other	
	Ga0302357_103245218	Other	
	Ga0302357_103245219	GH127	
	Ga0302357_103245220	Other	
	Ga0302357_103245221	Other	
	Ga0302357_103245222	Other	
	Ga0302357_103245223	GH43_10	3.2.1.55
	Ga0302357_103245224	GH95	3.2.1.x
	Ga0302357_103245225	GH127	
Ga0302357_1162480	Ga0302357_11624803	CE10	
	Ga0302357_11624804	CE10	
	Ga0302357_11624805	GH43_28_CBM32	
	Ga0302357_11624806	Other	
	Ga0302357_11624807	SusC	
	Ga0302357_11624808	SusD	
	Ga0302357_11624809	CE1_CBM48	
	Ga0302357_116248010	CE1_CBM48	3.1.1.73
	Ga0302357_116248011	CBM48_CE1	
	Ga0302357_116248012	GH127	
	Ga0302357_116248013	Other	
	Ga0302357_116248014	GH43_9	3.2.1.37
	Ga0302357_116248015	GH54	3.2.1.55
	Ga0302357_116248016	Other	
	Ga0302357_116248017	GH54	
	Ga0302357_116248018	CBM34_GH138	
	Ga0302357_116248019	GH51	3.2.1.55
	Ga0302357_116248020	GH97	3.2.1.3
	Ga0302357_116248021	GH3	3.2.1.21
	Ga0302357_116248022	GH29	
	Ga0302357_116248023	Other	
	Ga0302357_116248024	GH43_10	3.2.1.55
	Ga0302357_116248025	GH43_2_CBM6	
	Ga0302357_116248026	Other	

	Ga0302357_116248027	SusC	
	Ga0302357_116248028	SusD	
	Ga0302357_116248029	GH95	
	Ga0302357_116248030	GH95	
	Ga0302357_116248031	Other	
	Ga0302357_116248032	Other	
	Ga0302357_116248033	GH97	
	Ga0302357_116248034	Other	
	Ga0302357_116248035	GH95	
	Ga0302357_116248036	GH95	
	Ga0302357_116248037	GH95	
	Ga0302357_116248038	CBM6_GH3	3.2.1.37
	Ga0302357_116248039	GH95	
	Ga0302357_116248040	GH43_28_CBM32	
	Ga0302357_116248041	Other	
	Ga0302357_116248042	GH28	3.2.1.173
	Ga0302357_116248043	Other	
	Ga0302357_116248044	GH3	3.2.1.46

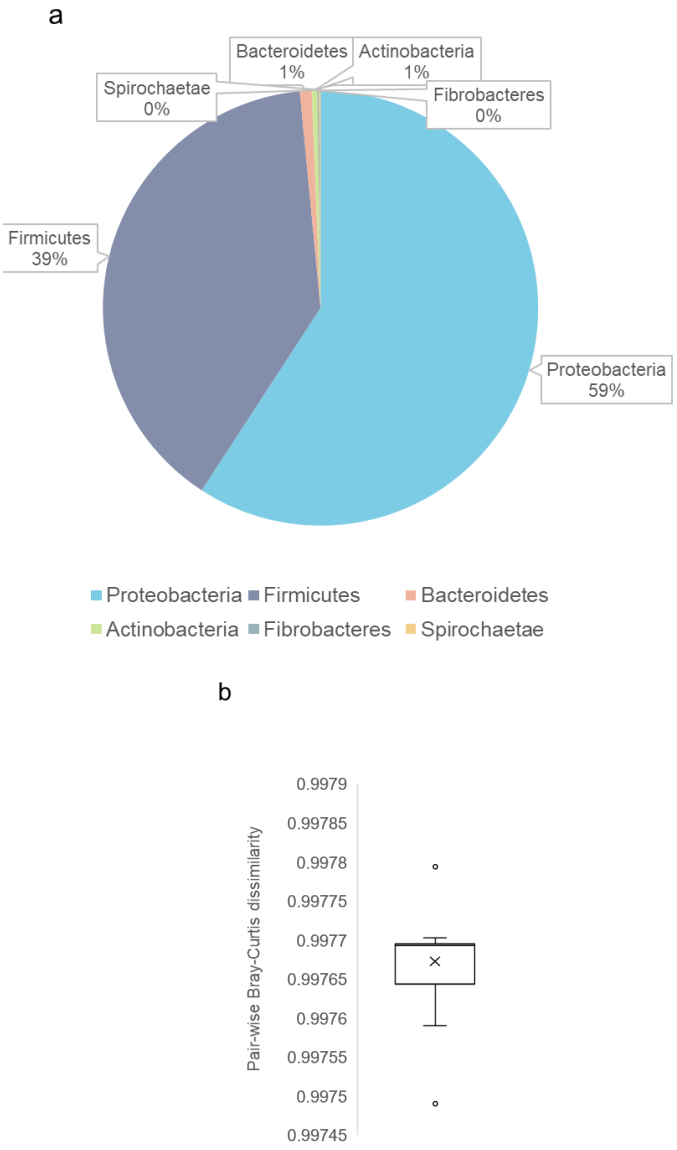
Annex 4: Supplementary material of Chapter 4



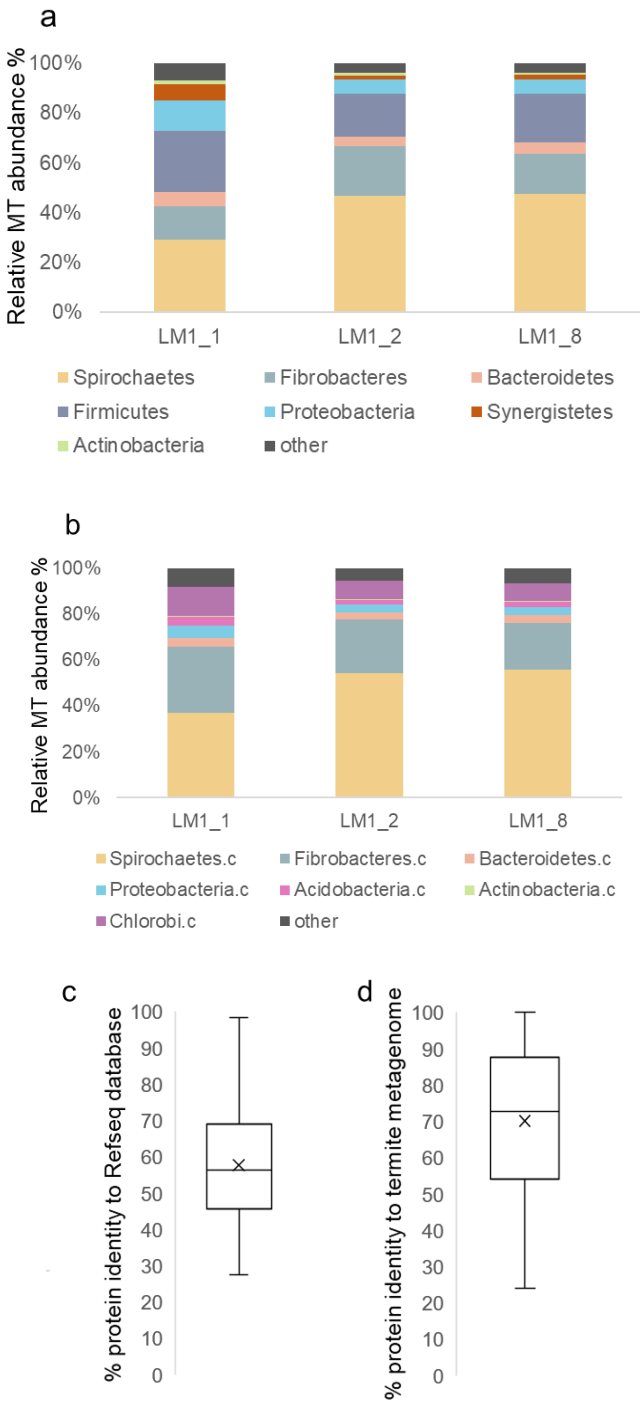
Supplementary Figure 1: Assay design of sampling and high-throughput sequencing characterisation of the termite gut lignocellulose digestion system. Termite hindguts from mature workers were sampled in regular monthly time intervals and nucleic acids were co-extracted. Colony LM2 was excluded from further analysis as it did not adapt to the laboratory fed miscanthus and diet. Control sample (fed original diet) is designated as LMx_1. Both the termite gut microbiome (bacterial community) and the termite gut epithelium were analysed.

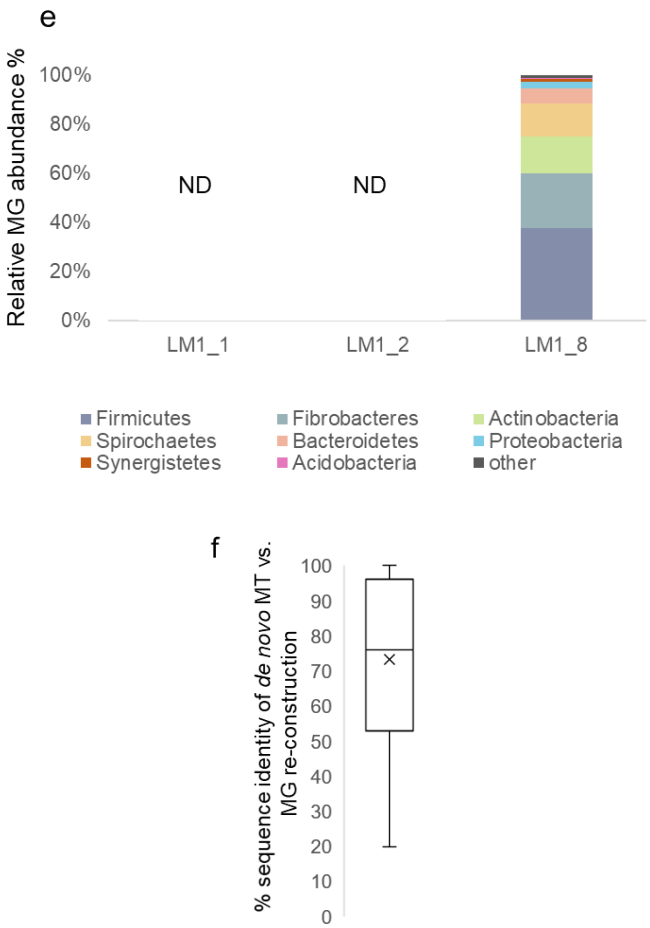


Supplementary Figure 2: Unrooted neighbor-joining tree of cytochrome oxidase II genes of the three termite species investigated in this study (LM1, LM2 and LM3) and their closest sequenced relatives, based on the homology search against the NCBI database. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. There were a total of 747 positions in the final dataset. Evolutionary analysis was conducted in MEGA X (1).

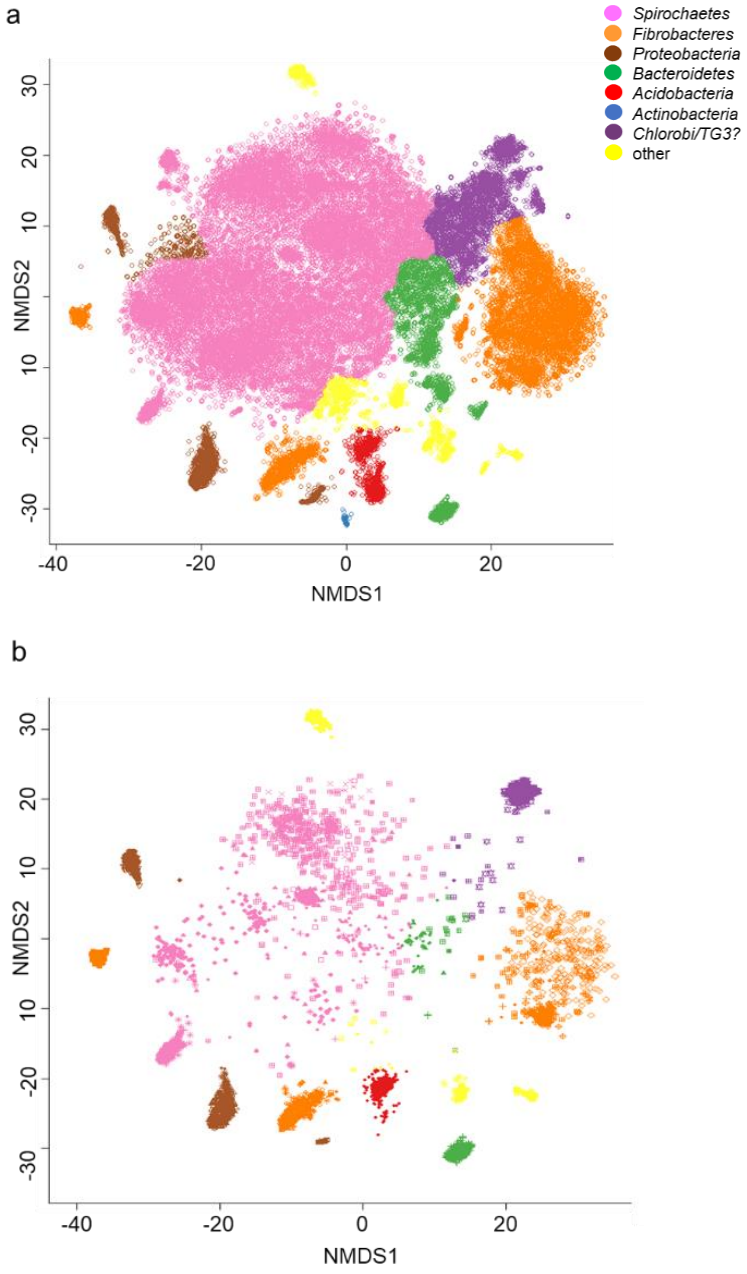


Supplementary Figure 3: Characterisation of the miscanthus straw associated bacteria. **(a)** Taxonomic distribution of the 16S rRNA reads to bacterial phyla. **(b)** Pair-wise Bray-Curtis dissimilarity between the miscanthus straw associated bacterial community and the miscanthus-adapted microbiome (bacterial community in the termite gut fed with miscanthus diet). The lowest calculated pair-wise Bray-Curtis dissimilarity distance was above 0.997 (on the scale from 0 to 1, where 0 means that two communities are identical and 1 means that they are maximally different), and none of the miscanthus straw associated bacteria was enriched in the termite gut microbiome. Box represents the interquartile range and error bars show the 95% confidence intervals.


a
Relative MT abundance %
100%
80%
60%
40%
20%
0%
LM1_1
LM1_2
LM1_8
Spirochaetes
Fibrobacteres
Bacteroidetes
Firmicutes
Proteobacteria
Synergistetes
Actinobacteria
other
b
Relative MT abundance %
100%
80%
60%
40%
20%
0%
LM1_1
LM1_2
LM1_8
Spirochaetes.c
Fibrobacteres.c
Bacteroidetes.c
Proteobacteria.c
Acidobacteria.c
Actinobacteria.c
Chlorobi.c
other
c
% protein identity to Refseq database
100
90
80
70
60
50
40
30
20
10
0
d
% protein identity to termite metagenome
100
90
80
70
60
50
40
30
20
10
0

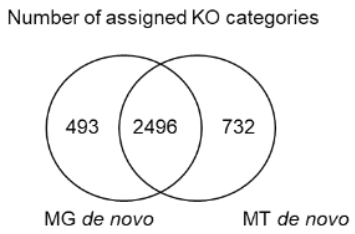


Supplementary Figure 4: **(a)** Database-dependent (IMG/MER; (2)) taxonomic assignment of reconstructed *de novo* metatranscriptomic (MT) gene transcripts for the studied termite gut microbiomes. **(b)** Taxonomic reclassification of the *de novo* MT reconstructed gene transcripts for the studied termite gut microbiomes based on their sequence homology to the *de novo* MG reconstructed contigs and bacterial MAGs. **(c, d)** Sequence similarity comparison of the *de novo* reconstructed gene transcripts for the studied termite gut microbiomes against the NCBI nucleotides database **(c)** and a custom *Nasutitermes* spp database **(d**; (3)). **(e)** Database-dependent (IMG/M; (2)) taxonomic assignment of reconstructed *de novo* metagenomic (MG) genes for the studied termite gut microbiome. **(f)** Comparison of the de novo MG and MT reconstructions, based on the sequence similarity of the reconstructed genes (MG) versus gene transcripts (MT). ND – sequencing not done. **(c, d, f)** Boxes represent the interquartile range and error bars show the 95% confidence intervals.

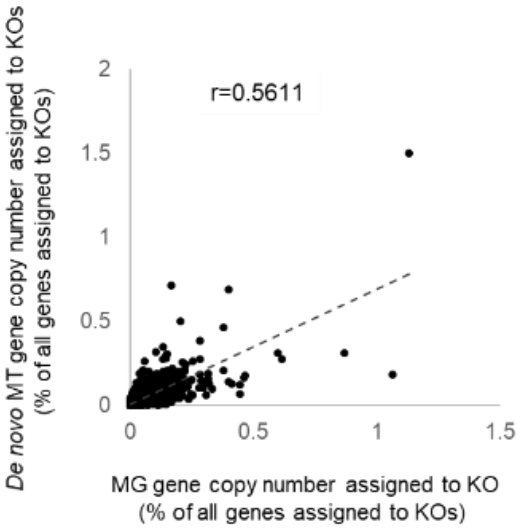


Supplementary Figure 5: Non-metric multidimensional (NMDS) scaling plot illustrating similarities of **(a)** Binning results of the *de novo* reconstructed metagenomic contigs to phylum-level bins and taxonomic bin assignment with PhyloPhlan (4) and **(b)** Reconstruction of species-level MAGs (metagenome assembled genomes) and their phylogenetic classification. Points represent (a) bins and (b) MAGs. The closer are two points, the higher is the similarity between them.

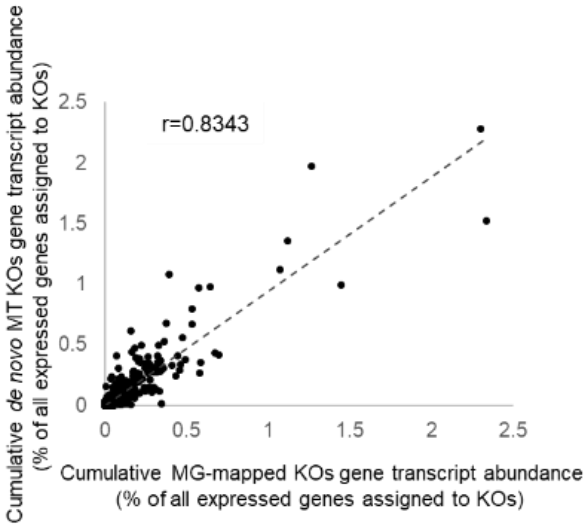
a


Number of assigned KO categories
493
2496
732
MG *de novo*
MT *de novo*

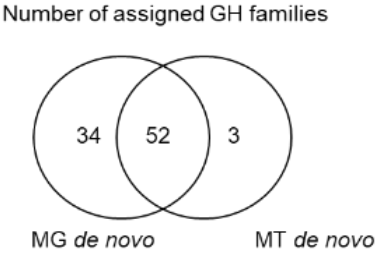
b


r=0.5611
De novo MT gene copy number assigned to KOs
(% of all genes assigned to KOs)
MG gene copy number assigned to KO
(% of all genes assigned to KOs)
0
0.5
1
1.5
2

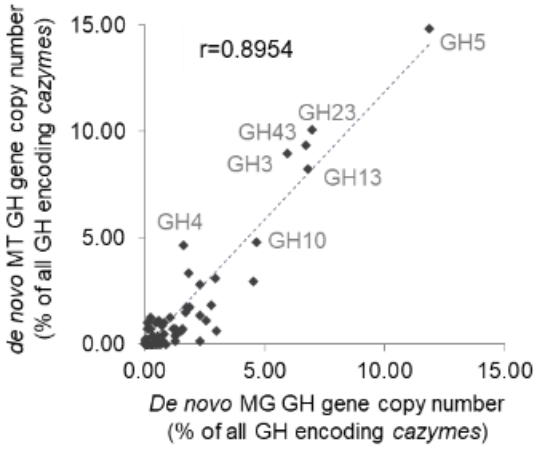
c


r=0.8343
Cumulative *de novo* MT KOs gene transcript abundance
(% of all expressed genes assigned to KOs)
Cumulative MG-mapped KOs gene transcript abundance
(% of all expressed genes assigned to KOs)
0
0.5
1
1.5
2
2.5

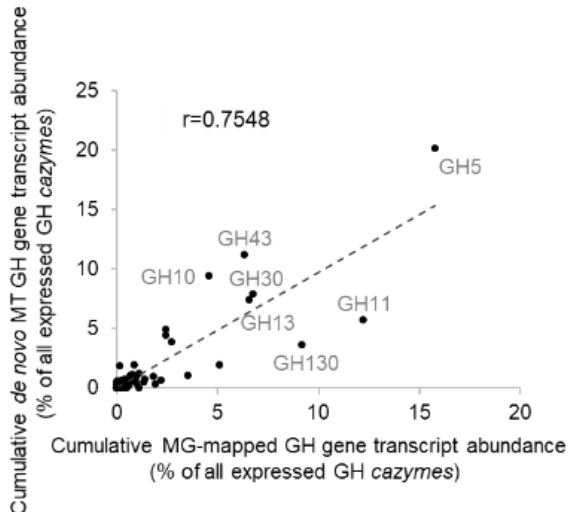
d

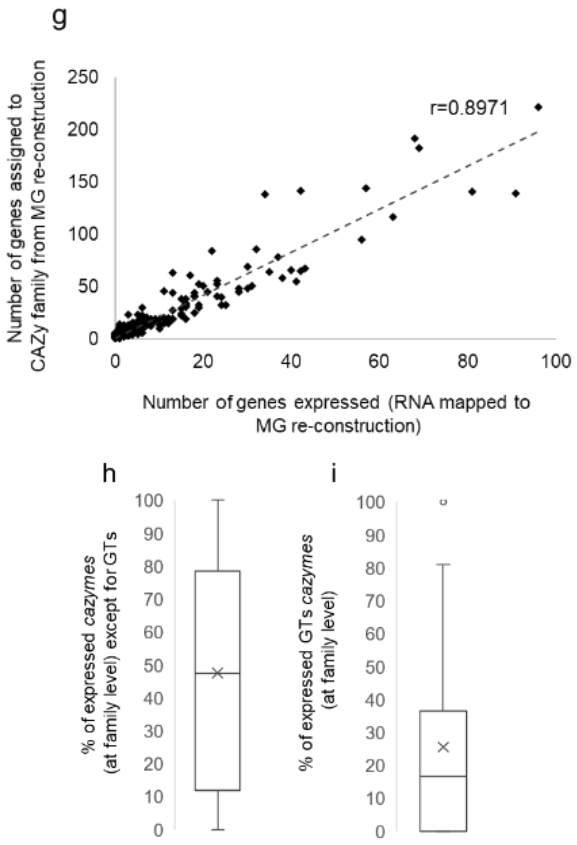

Number of assigned GH families
34
52
3
MG *de novo*
MT *de novo*

e

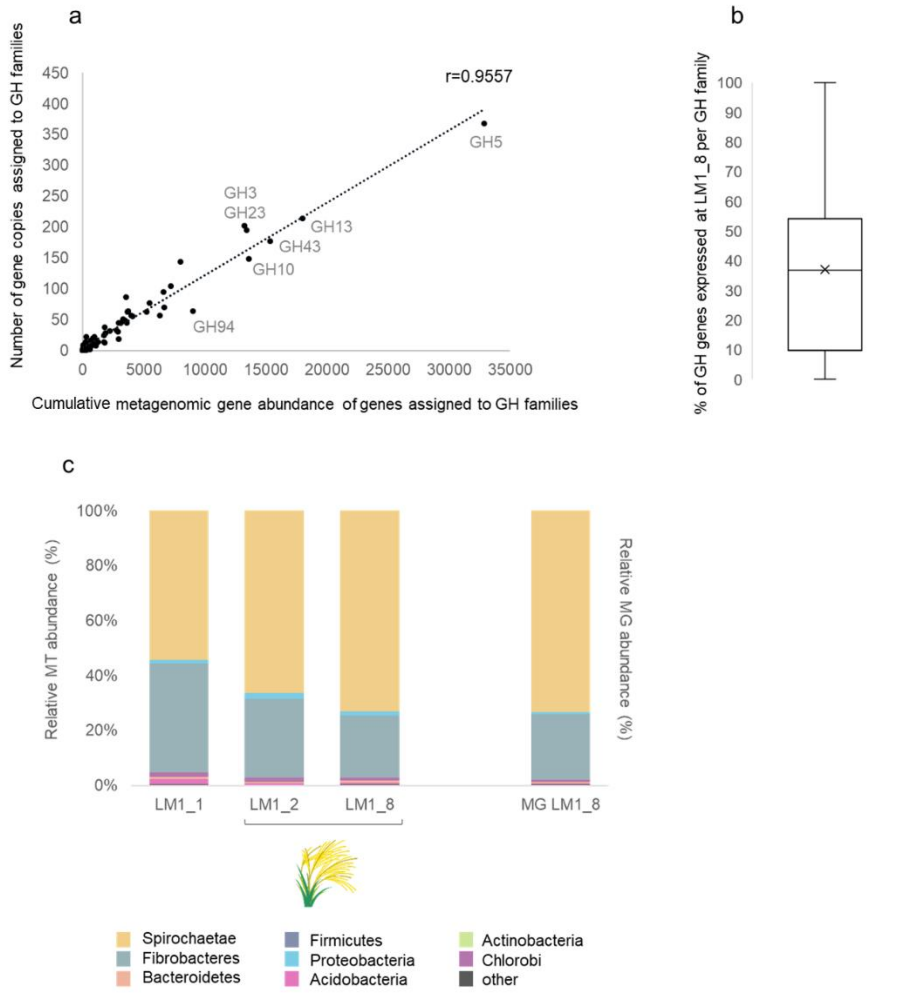

r=0.8954
GH5
GH23
GH43
GH3
GH13
GH4
GH10
15.00
10.00
5.00
0.00
0.00
5.00
10.00
15.00
de novo MT GH gene copy number
(% of all GH encoding *cazymes*)
De novo MG GH gene copy number
(% of all GH encoding *cazymes*)

f

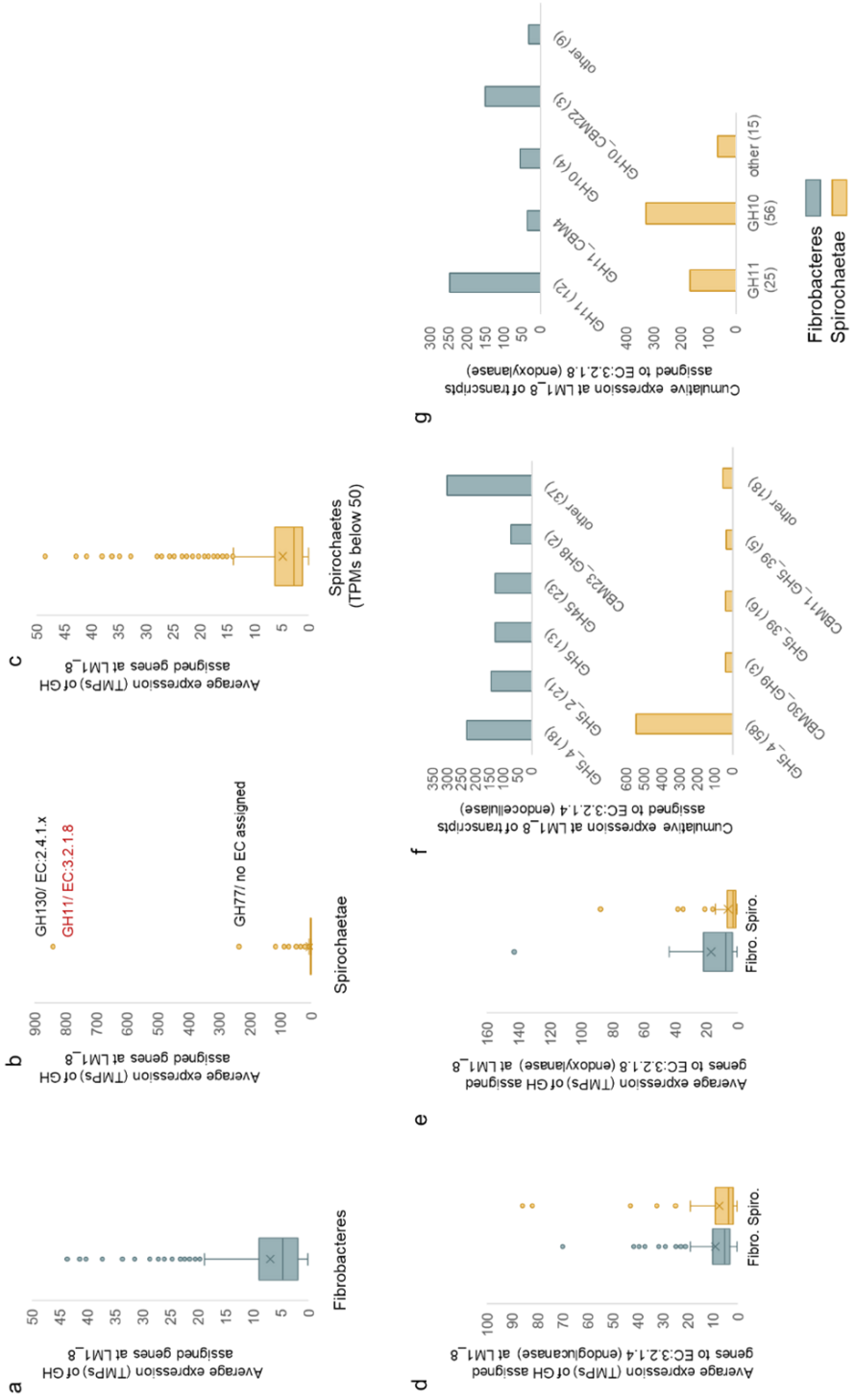

r=0.7548
GH5
GH43
GH10
GH30
GH11
GH13
GH130
25
20
15
10
5
0
0
5
10
15
20
Cumulative *de novo* MT GH gene transcript abundance
(% of all expressed GH *cazymes*)
Cumulative MG-mapped GH gene transcript abundance
(% of all expressed GH *cazymes*)



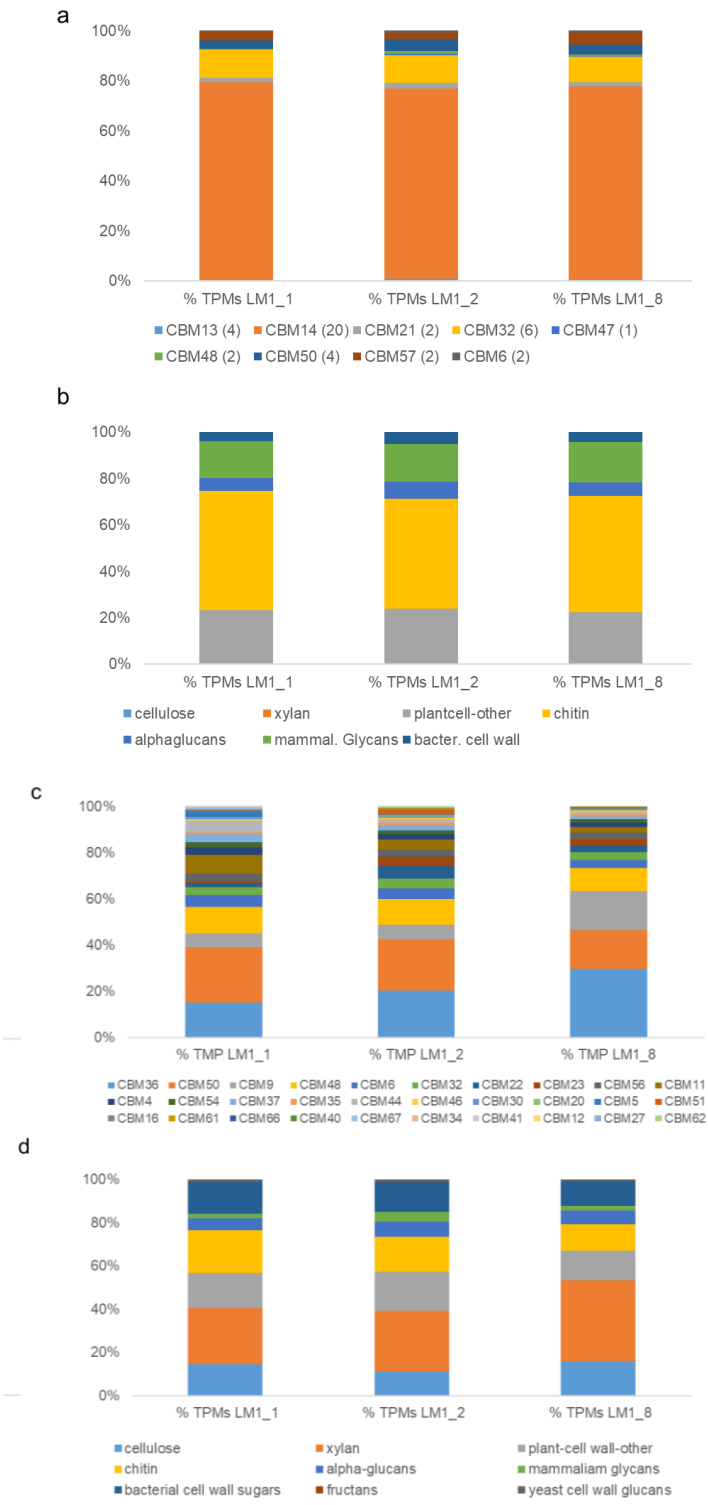
Supplementary Figure 6: Comparison of the *de novo* metagenomic (MG) and metatranscriptomic (MT) reconstructions for the studied termite gut microbiomes. **(a)** Venn diagram showing unique and shared KOs between the two datasets. **(b)** Comparison of the number of reconstructed genes (MG) and mapped gene transcripts (MT) assigned to the same KO category for sample LM1_8. Pearson coefficient of correlation is displayed on the graph. **(c)** Comparison of the cumulative gene (MG) and gene transcript (MT) abundance assigned to the same KO category. Pearson coefficient of correlation is displayed on the graph. **(d)** Venn diagram showing unique and shared glycoside hydrolase (GH) families between the two datasets. **(e)** Comparison of the absolute number (i.e. gene copy number on the graph) of reconstructed genes (MG) and gene transcripts (MT) assigned to the same GH family. Pearson coefficient of correlation is displayed on the graph. **(f)** Comparison of the cumulative gene (MG) and mapped gene transcripts (MT) abundance assigned to the same GH family for sample LN1_8. Pearson coefficient of correlation is displayed on the graph. **(g)** Comparison of the number of assigned genes (MG) to GH families and expressed genes (MT) assigned to the same GH family for sample LN1_8. Pearson coefficient of correlation is displayed on the graph. **(h, i)** Box plots representation with the median, first and third quartiles displayed of the average gene number expression per CAZyme family, with a separate focus on glycosyl transferases GTs (i). **(h,i)** Boxes represent the interquartile range and error bars show the 95% confidence intervals.

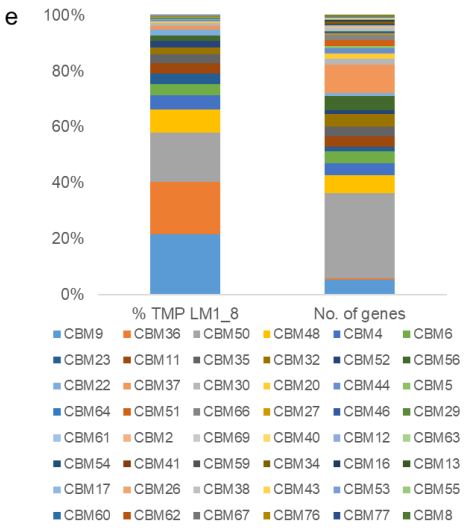


Supplementary Figure 7: Characterisation of bacterial genes assigned to the different glycoside hydrolases (GHs) families for the *de novo* metagenomic (MG) reconstruction. **(a)** Correlation between the absolute number of genes (i.e. number of gene copies on the graph) assigned to a GH family and their cumulative MG abundance. In the case of the GH11 family, the highly abundant and partially reconstructed gene outliers were not displayed on the graph. Pearson coefficient of correlation is displayed on the graph. **(b)** Number of the *de novo* MG reconstructed genes assigned to the different GH families that were expressed at the time point LM1_8; given per family (RNA-seq). Box represents the interquartile range and error bars show the 95% confidence intervals. **(c)** Database-independent classification of the *de novo* reconstructed genes (MG) and gene transcripts (MT) to the phylum level, and based on the MG contig binning and bin taxonomic annotation with PhyloPhlan (4).

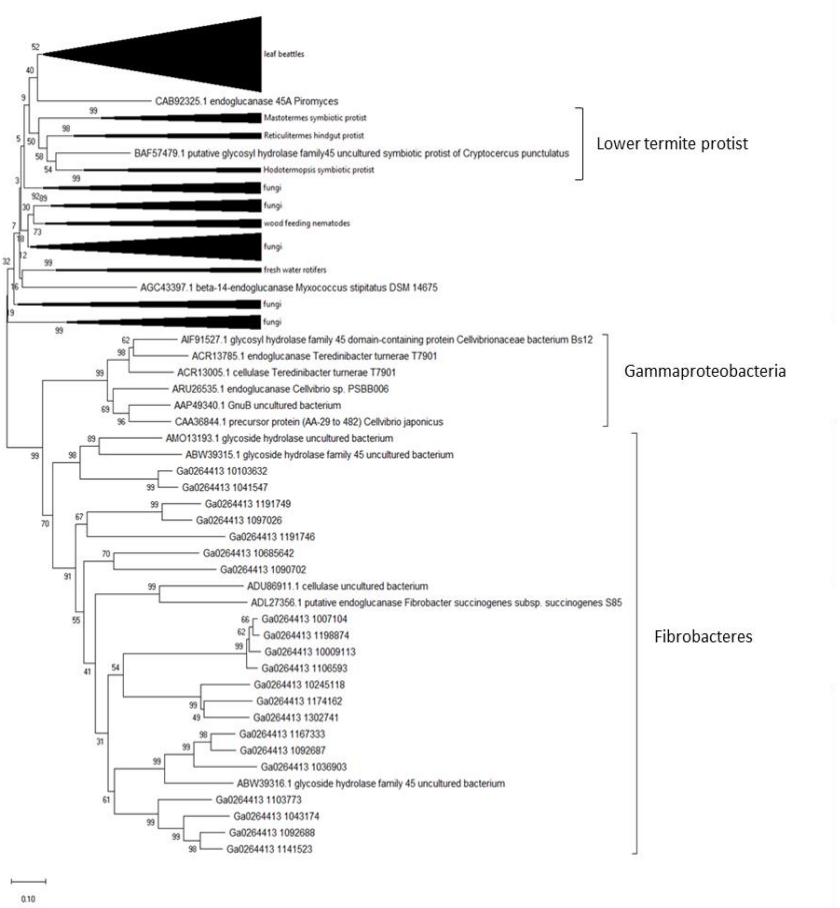

a
Average expression (TMPs) of GH assigned genes at LM1_8
Fibrobacteres
b
Average expression (TMPs) of GH assigned genes at LM1_8
GH130/ EC:2.4.1.x
GH11/ EC:3.2.1.8
GH77/ no EC assigned
Spirochaetae
c
Average expression (TMPs) of GH assigned genes at LM1_8
Spirochaetes (TPMs below 50)
d
Average expression (TMPs) of GH assigned genes to EC:3.2.1.4 (endoglucanase) at LM1_8
Fibro. Spiro.
e
Average expression (TMPs) of GH assigned genes to EC:3.2.1.8 (endoxylanase) at LM1_8
Fibro. Spiro.
f
Cumulative expression at LM1_8 of transcripts assigned to EC:3.2.1.4 (endocellulase)
GH5_4 (18)
GH5_2 (21)
GH5 (13)
GH45 (23)
CBM23_GH8 (2)
other (37)
GH5_4 (58)
CBM30_GH9 (3)
GH5_39 (16)
CBM11_GH5_39 (5)
other (18)
g
Cumulative expression at LM1_8 of transcripts assigned to EC:3.2.1.8 (endoxylanase)
GH11 (12)
GH11_CBM4
GH10 (4)
GH10_CBM22 (3)
other (9)
GH11 (25)
GH10 (56)
other (15)
Fibrobacteres
Spirochaetae

Supplementary Figure 8: Phylum level (*Spirochaetae* and *Fibrobacteres*) characterisation of the termite gut bacterial glycoside hydrolase (GH) coding genes and their expression profiles. Average metatranscriptomic (MT) abundance of gene transcripts assigned to the different GH families for **(a)** *Fibrobacteres* and *Spirochaetae* **(b, c)**. Average MT abundance of GH gene transcripts functionally assigned to endoglucanases **(d)** and endoxylanases **(e)**. **(d, e)** Boxes represent the interquartile range and error bars show the 95% confidence intervals. Cumulative MT abundance of GH gene transcripts functionally assigned to endoglucanases **(f)** and endoxylanases **(g)**. Distribution of gene transcripts between the different GH families is shown as bar charts. Number of reconstructed gene transcripts is shown in brackets. Gene transcripts outliers (highly abundant but partially reconstructed gene transcripts) were removed from panels c, e and g.

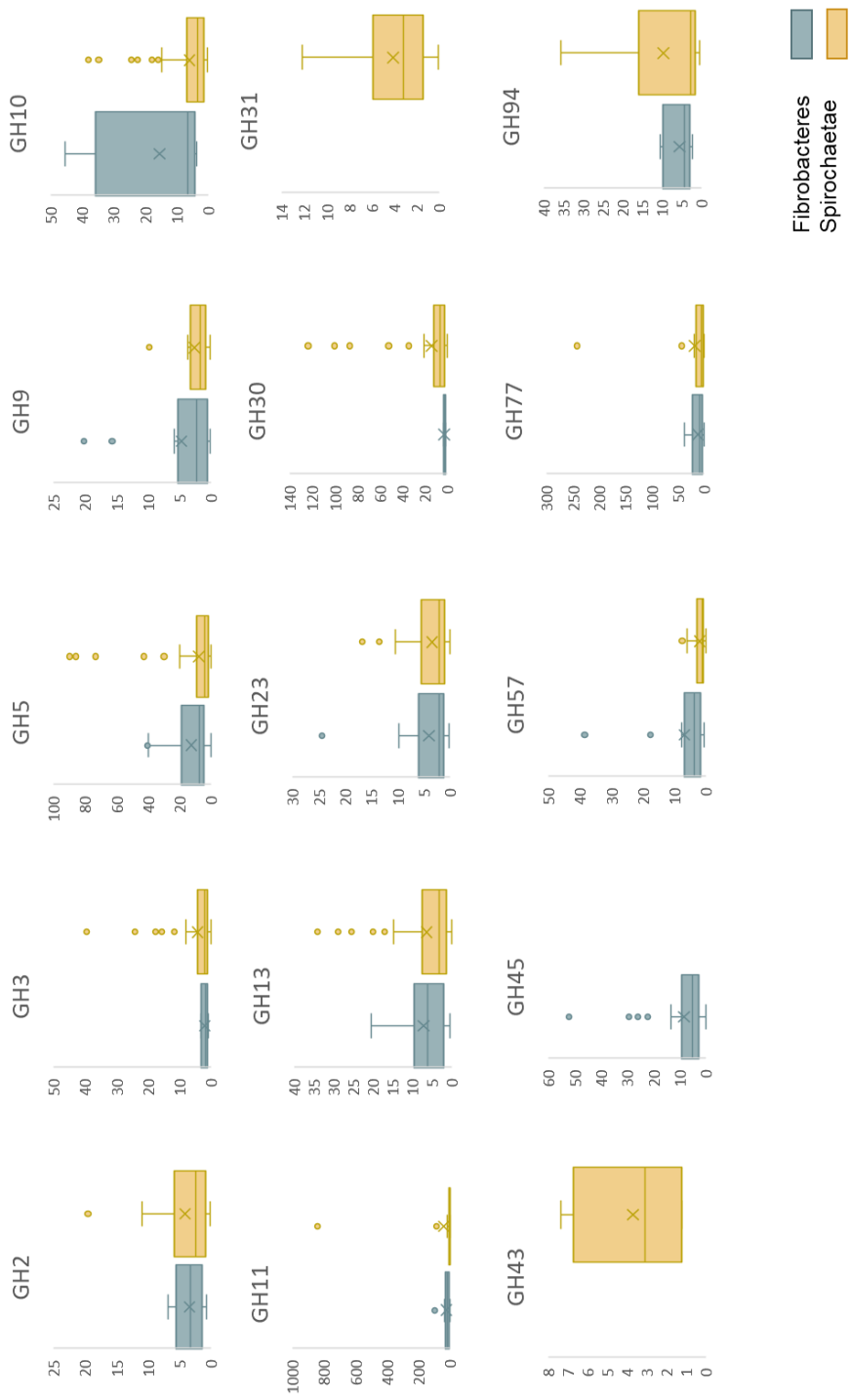

a
100%
80%
60%
40%
20%
0%
% TPMs LM1_1
% TPMs LM1_2
% TPMs LM1_8
CBM13 (4) CBM14 (20) CBM21 (2) CBM32 (6) CBM47 (1)
CBM48 (2) CBM50 (4) CBM57 (2) CBM6 (2)
b
100%
80%
60%
40%
20%
0%
% TPMs LM1_1
% TPMs LM1_2
% TPMs LM1_8
cellulose xylan plantcell-other chitin
alphaglucans mammal. Glycans bacter. cell wall
c
100%
80%
60%
40%
20%
0%
% TMP LM1_1
% TMP LM1_2
% TMP LM1_8
CBM36 CBM50 CBM9 CBM48 CBM6 CBM32 CBM22 CBM23 CBM56 CBM11
CBM4 CBM54 CBM37 CBM35 CBM44 CBM46 CBM30 CBM20 CBM5 CBM51
CBM16 CBM61 CBM66 CBM40 CBM67 CBM34 CBM41 CBM12 CBM27 CBM62
d
100%
80%
60%
40%
20%
0%
% TPMs LM1_1
% TPMs LM1_2
% TPMs LM1_8
cellulose xylan plant-cell wall-other
chitin alpha-glucans mammaliam glycans
bacterial cell wall sugars fructans yeast cell wall glucans



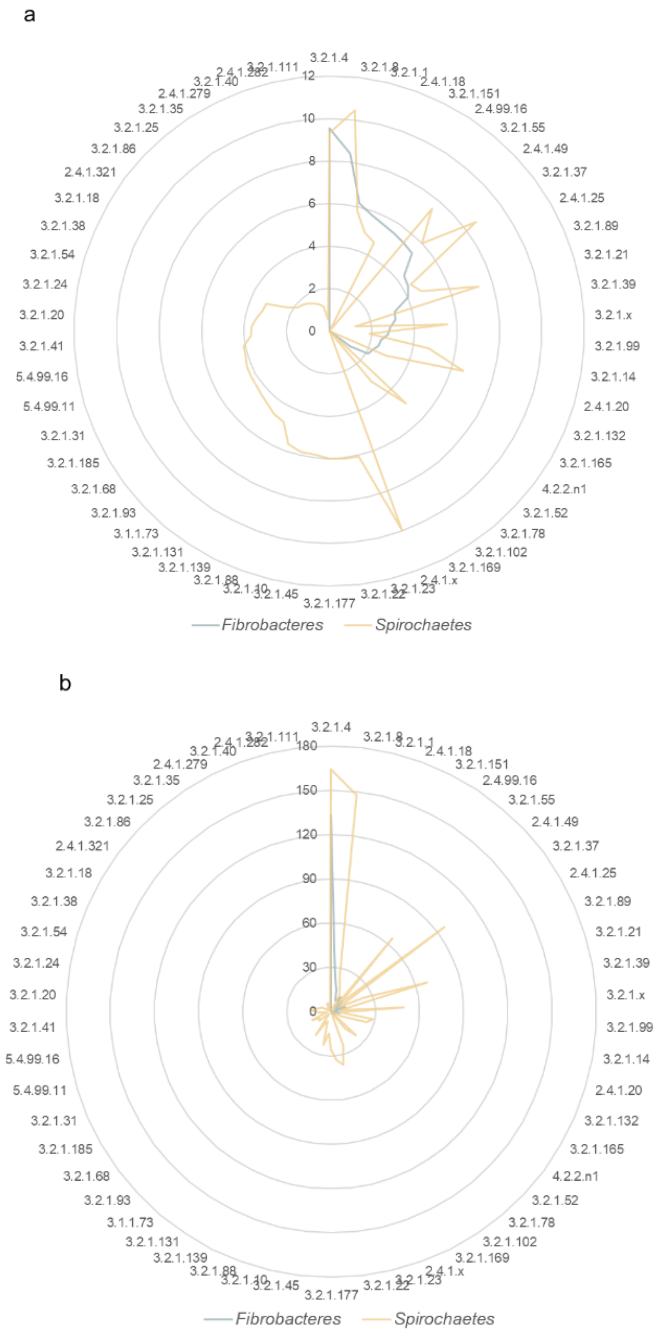
Supplementary Figure 9: Characterisation of termite and bacterial genes (*de novo* metagenomics (MG)) and gene transcripts (*de novo* metatranscriptomics (MT)) assigned to the different carbohydrate binding module (CBM) families. **(a)** Distribution of reconstructed gene transcripts of termite origin to CBM families. Total number of genes is given in brackets. **(b)** Substrate specificity of reconstructed CBM encoding genes of termite origin, based on known substrate specificities of different CBM families. **(c)** Distribution of reconstructed gene transcripts (*de novo* MT) of bacterial origin to CBM families. **(d)** Substrate specificity of reconstructed CBM encoding genes of bacterial origin, based on known substrate specificities of different CBM families **(e)** Metagenomic abundance and gene number of the different CBM families, based on the annotation of the *de novo* reconstructed genes (*de novo* MG). For all graphs, the data is expressed as % of total CBM gene transcript abundance or gene count. Note that the colour code may change between the different panel in this figure.



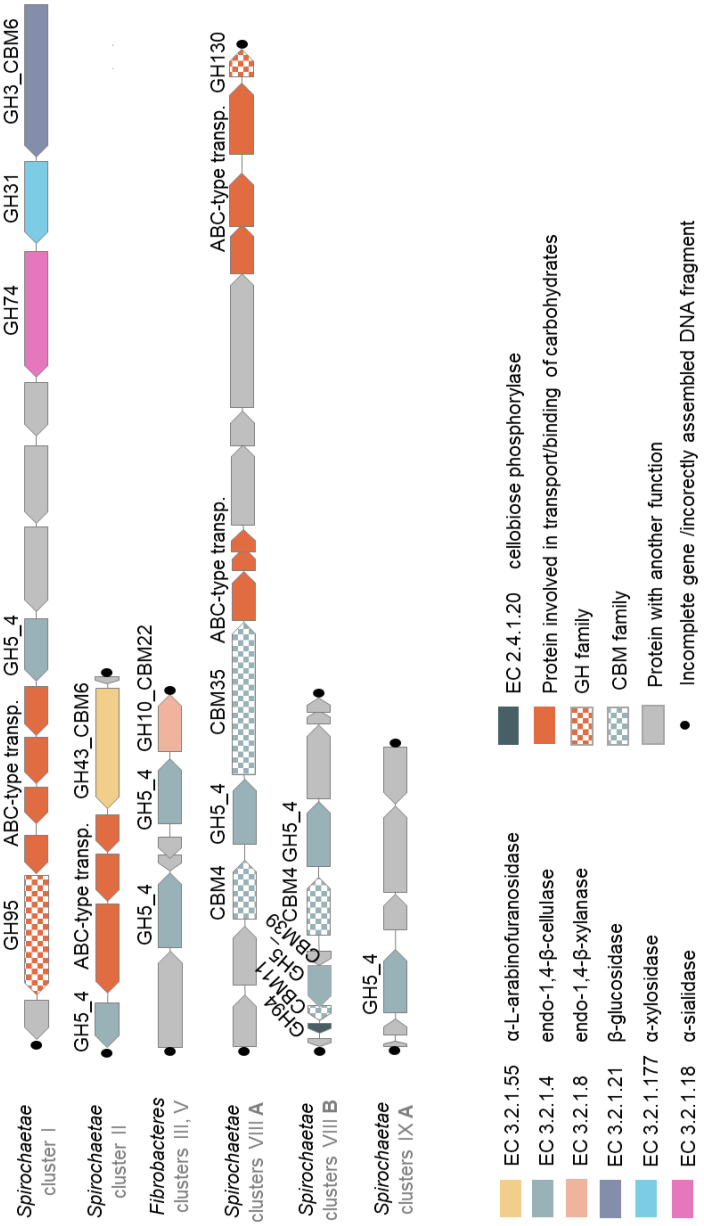
Supplementary Figure 10: Unrooted neighbor-joining tree of GH45 assigned genes, based on the *de novo* metagenomic (MG) reconstruction. All known bacterial and eukaryotic genes assigned to GH45 in the CAZY database (http://www.cazy.org) are included. Some short sequences were removed from final alignment to increase the number of final positions in the alignment that could have been compared. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. There were a total of 486 positions in the final dataset. Evolutionary analysis was conducted in MEGA X (1).


GH10
GH31
GH94
GH9
GH30
GH77
GH5
GH23
GH57
GH3
GH13
GH45
GH2
GH11
GH43
Fibrobacteres
Spirochaetae

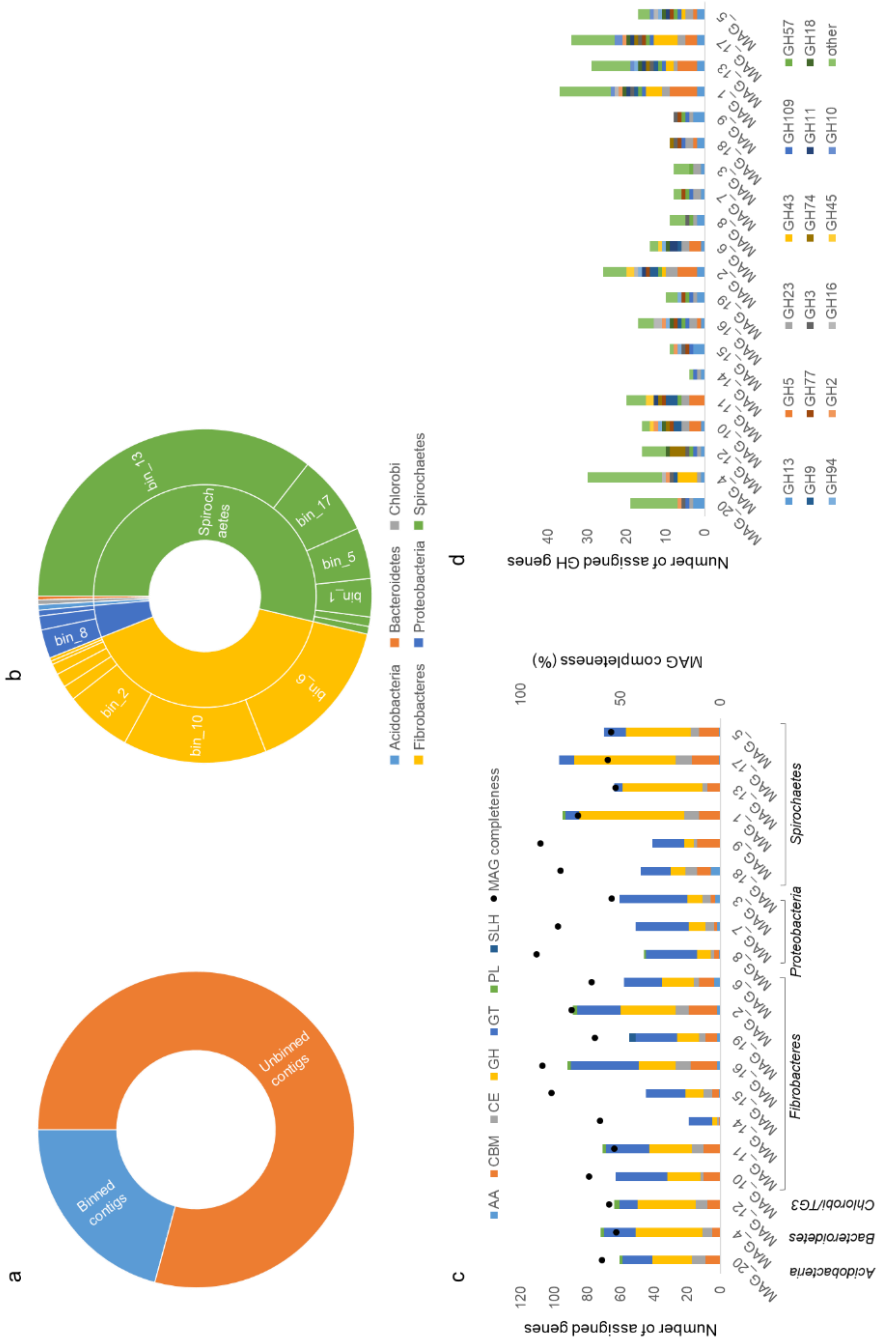
Supplementary Figure 11: Average gene expression (TMP) for *Spirochaetae* and *Fibrobacteres* of assigned glycoside hydrolase (GH) genes per GH family (RNA-seq results) at LM1_8. Box plots representation with the median, first and third quartiles is displayed. Outliers (highly expressed genes; in some case representing only partially reconstructed genes) are indicated on the different panels with a dot. Boxes represent the interquartile range and error bars show the 95% confidence intervals.



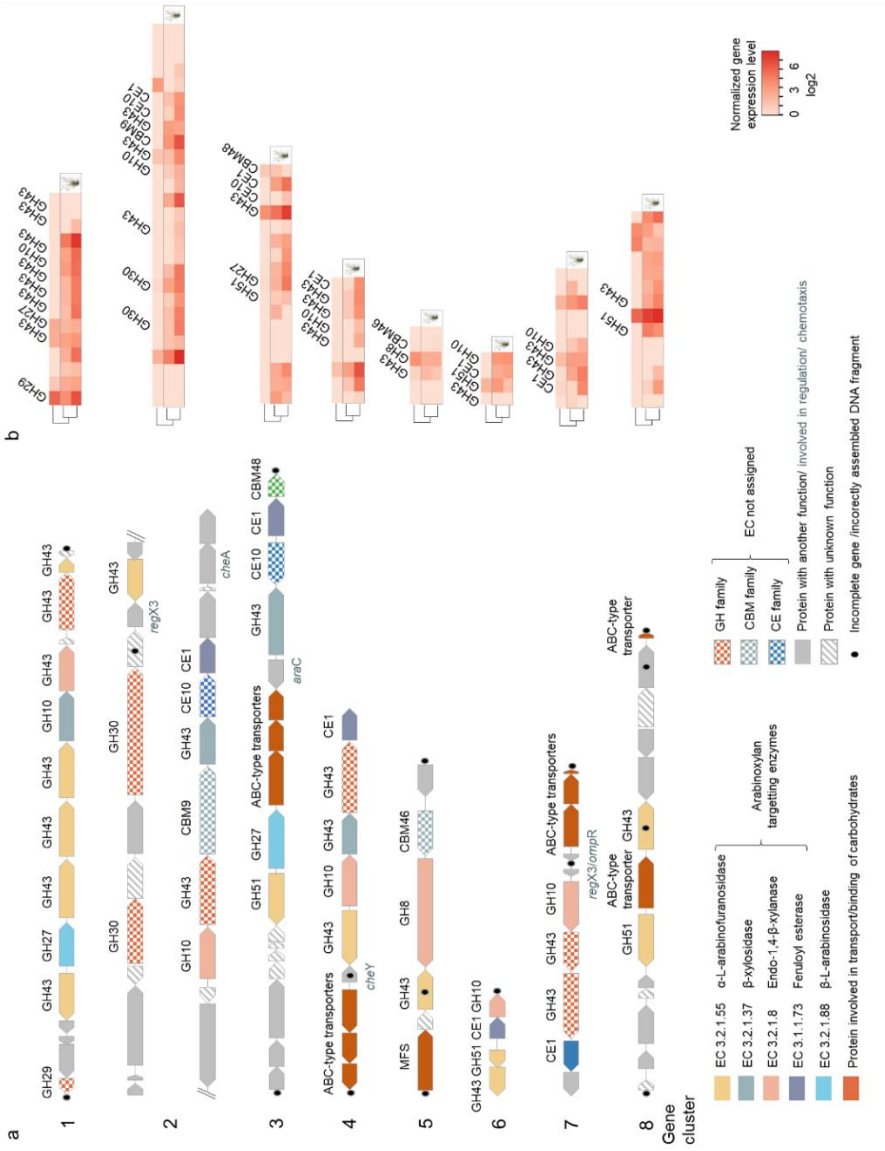
Supplementary Figure 12: Distribution of all glycoside hydrolase (GH) -assigned genes of *Fibrobacteres* and *Spirochaetae* origin to different EC categories. **(a)** Relative transcriptional abundance (expressed as TPMs, log2 transformed) of reconstructed gene transcripts (*de novo* metatrancriptomics (MT)) at LM1_8. **(b)** Absolute number of genes.



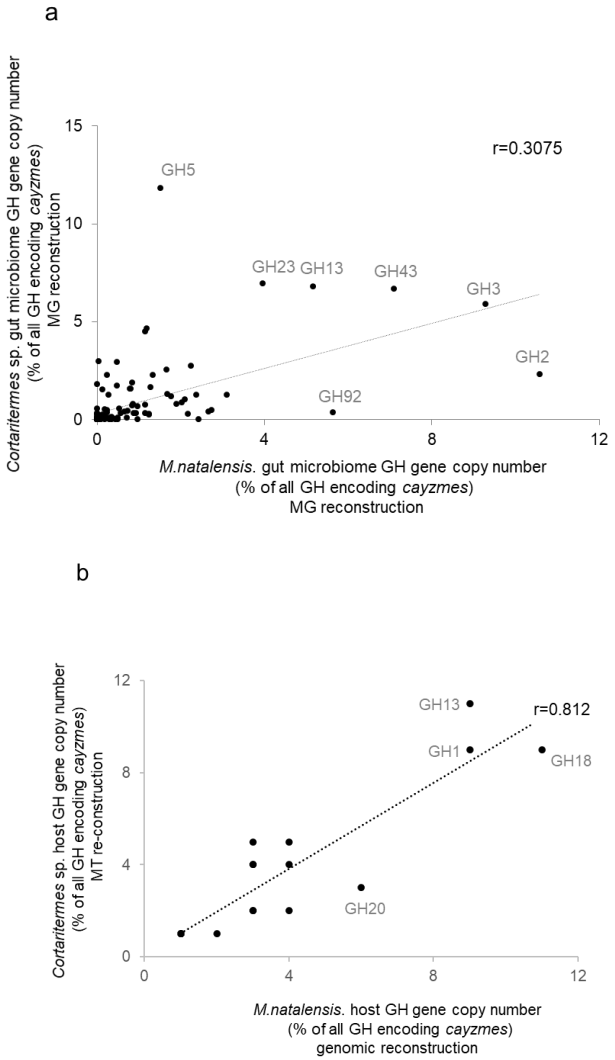
Supplementary Figure 13: Schematic representation of gene organisation within putative CAZymes genes loci. Genomic reconstructions are partial, therefore the CAZymes clusters may be fragmented (incomplete). Clusters designation (I to IX) refers to the Fig. 5a in the Chapter 4 of the manuscript.


a
Unbinned contigs
Binned contigs
b
Spirochaetes
Proteobacteria
Bacteroidetes
Chlorobi
Fibrobacteres
Acidobacteria
bin_17
bin_5
bin_1
bin_6
bin_10
bin_2
bin_8
bin_13
c
Number of assigned genes
MAG completeness (%)
AA
CBM
CE
GH
GT
PL
SLH
MAG completeness
Acidobacteria
Bacteroidetes
Chloroflexi/TG3
Fibrobacteres
Proteobacteria
Spirochaetes
d
Number of assigned GH genes
GH13
GH9
GH94
GH5
GH77
GH2
GH23
GH3
GH16
GH43
GH74
GH45
GH109
GH11
GH10
GH57
GH18
other

Supplementary Figure 14: Characterisation of reconstructed metagenome assembled genomes (MAGs). **(a)** Proportion of reconstructed contigs that were further binned to MAGs, out of the whole *de novo* metagenomic (MG) reconstruction for sample LM1_8. **(b)** Metagenomic abundance of reconstructed MAGs (proportion of all MAGs) based on the average abundance of reconstructed contigs binned into specific MAGs. MAGs are coloured according to their phylum-level taxonomic assignment. **(c)** CAZymes content of the reconstructed MAGs. **(d)** Dominant glycoside hydrolases (GHs) present in the reconstructed MAGs.

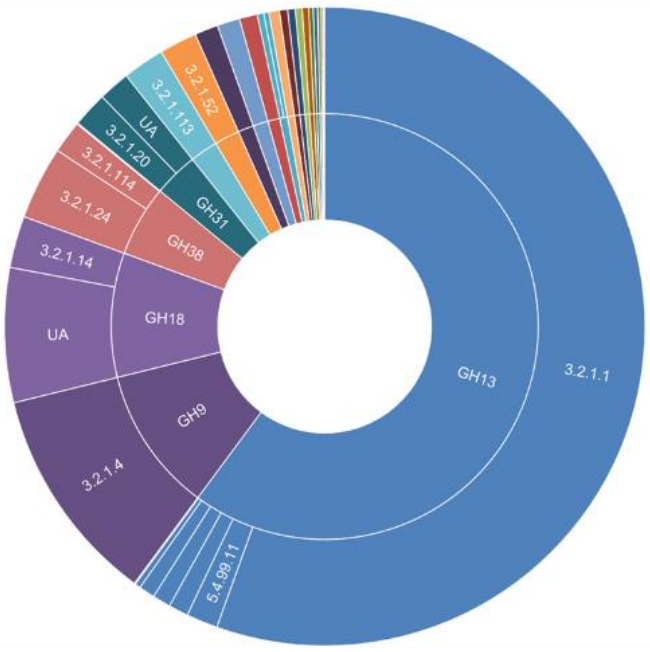


Supplementary Figure 15: Schematic representation of gene organisation within putative arabinoxylan-targeting CAZymes clusters of *Spirochaetae* origin **(a)**. Genomic reconstructions are partial, therefore the CAZymes clusters may be fragmented. **(b)** Gene expression levels before and under miscanthus diet for genes indicated in panel a.



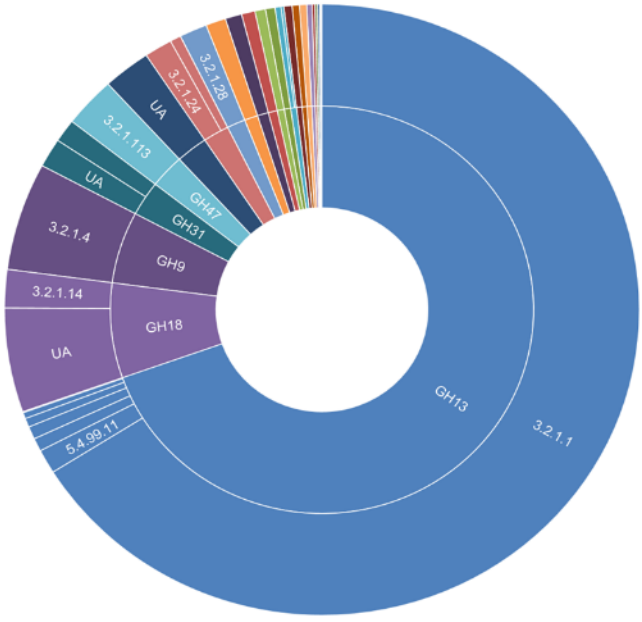
Supplementary Figure 16: Comparison of the glycoside hydrolase (GH) content of the studied *Cortaritermes* gut system with a previously studied M. *natalensis* (5). For M. *natalensis* GH content was derived from dbCAN2 analysis (Supplementary Data 9). **(a)** Comparison of the absolute number of GH gene between the gut microbiomes of the two studied species. **(b)** Comparison of the absolute number of GH gene between the two studied termite species. Remark : GH gene copy number refers to absolute number of GH

a

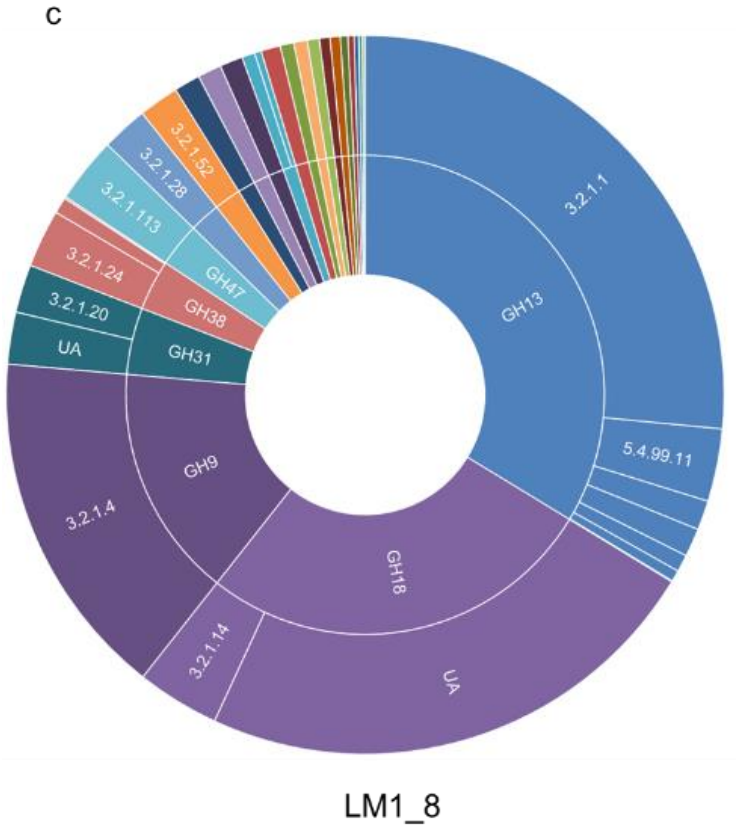

GH13
3.2.1.1
5.4.99.11
GH9
3.2.1.4
UA
GH18
3.2.1.14
3.2.1.24
GH38
3.2.1.114
3.2.1.20
UA
GH31
3.2.1.113
3.2.1.52

LM1_1

b


GH13
3.2.1.1
5.4.99.11
UA
GH18
3.2.1.14
3.2.1.4
GH9
GH31
GH47
UA
3.2.1.113
UA
3.2.1.24
3.2.1.28

LM1_2



Supplementary Figure 17: Distribution of glycoside hydrolase (GH) -assigned gene transcripts of *Cortaritermes* sp. origin to different EC categories. **(a)** Relative transcriptional abundance (expressed as TPMs, log2 transformed) of reconstructed gene transcripts (*de novo* metatranscriptomics (MT)) at LM1_1; **(b)** at LM1_2; and **(c)** at LM1_8.

Supplementary Data 1

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM3_ESM.xlsx

Supplementary Data 2

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM4_ESM.xlsx

Supplementary Data 3

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM5_ESM.xlsx

Supplementary Data 4

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM6_ESM.xlsx

Supplementary Data 5

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM7_ESM.xlsx

Supplementary Data 6

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM8_ESM.xlsx

Supplementary Data 7

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM9_ESM.xlsx

Supplementary Data 8

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM10_ESM.xlsx

Supplementary Data 9

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM11_ESM.xlsx

Supplementary Data 10

https://static-content.springer.com/esm/art%3A10.1038%2Fs42003-020-1004-3/MediaObjects/42003_2020_1004_MOESM12_ESM.xlsx

Supplementary references:

1. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
2. I. M. A. Chen, *et al.*, IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
3. K. Rossmassler, *et al.*, Metagenomic analysis of the microbiota in the highly compartmented hindguts of six wood- or soil-feeding higher termites. *Microbiome* **3**, 56 (2015).
4. N. Segata, D. Börnigen, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4** (2013).
5. M. Poulsen, *et al.*, Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 14500–14505 (2014).