Characterization of three bacterial glycoside hydrolase family 9 endoglucanases with different modular architectures isolated from a compost metagenome

Laure Aymé, Agnès Hébert, Bernard Henrissat, Vincent Lombard, Nathalie Franche, Stéphanie Perret, Etienne Jourdier, Senta Heiss-Blanquet

Characterization of three bacterial glycoside hydrolase family 9 endoglucanases with different modular architectures isolated from a compost metagenome

Laure Aymé[a], Agnès Hébert[a], Bernard Henrissat[b,c,d], Vincent Lombard[b,c], Nathalie Franche[e], Stéphanie Perret[e], Etienne Jourdier[a] and Senta Heiss-Blanquet[a*]

[a]IFP Energies Nouvelles, 1 - 4 avenue du Bois-Préau, 92852 Rueil-Malmaison, France

[b]Architecture et Fonction des Macromolécules Biologiques (AFMB), CNRS, 163 avenue de Luminy, 13288 Aix Marseille Université, Marseille, France

[c]INRAE, USC1408 Architecture et Fonction des Macromolécules Biologiques (AFMB), 163 avenue de Luminy, 13288 Marseille, France

[d]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[e]Aix Marseille Université, CNRS, LCB, 31 Chemin Joseph Aiguier, 13009 Marseille, France

*Corresponding author:

Senta Heiss-Blanquet, IFP Energies Nouvelles, 1 et 4, avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France

+33 1 47 52 72 56, senta.blanquet@ifpen.fr, ORCID 0000-0001-8533-3274

Abstract

Background: Environmental bacteria express a wide diversity of glycoside hydrolases (GH). Screening and characterization of GH from metagenomic sources provides an insight into biomass degradation strategies of non-cultivated prokaryotes.

Methods: In the present report, we screened a compost metagenome for lignocellulolytic activities and identified six genes encoding enzymes belonging to family GH9 (GH9a-f). Three of these enzymes (GH9b, GH9d and GH9e) were successfully expressed and characterized.

Results: A phylogenetic analysis of the catalytic domain of pro- and eukaryotic GH9 enzymes suggested the existence of two major subgroups. Bacterial GH9s displayed a wide variety of modular architectures and those harboring an N-terminal Ig-like domain, such as GH9b and GH9d, segregated from the remainder. We purified and characterized GH9 endoglucanases from both subgroups and examined their stabilities, substrate specificities and product profiles. GH9e exhibited an original hydrolysis pattern, liberating an elevated proportion of oligosaccharides longer than cellobiose. All of the enzymes exhibited processive behavior and a synergistic action on crystalline cellulose. Synergy was also evidenced between GH9d and a GH48 enzyme identified from the same metagenome.

Conclusions: The characterized GH9 enzymes displayed different modular architectures and distinct substrate and product profiles. The presence of a cellulose binding domain was shown to be necessary for binding and digestion of insoluble cellulosic substrates, but not for processivity.

General significance: The identification of six GH9 enzymes from a compost metagenome and the functional variety of three characterized members highlight the importance of this enzyme family in bacterial biomass deconstruction.

Keywords: metagenome, cellulose hydrolysis, endoglucanase, glycoside hydrolase

1. Introduction

Carbohydrates are the building blocks of the most structurally diverse class of biopolymers, such as α- and β-glucans, chitosan, xanthan, agar or pectins [1]. Nature has thus evolved a huge variety of enzymes able to synthesize, modify or degrade these polymers. Among them, glycoside hydrolases (GHs) (EC 3.2.1.x) catalyzing the hydrolysis of glycosidic bonds are of special interest as their activity is of major importance for the deconstruction of plant- and animal-derived materials such as starch, cellulose or chitin [2]. GHs are one of the five enzyme classes which are classified further in the comprehensive Carbohydrate-Active enZymes (CAZy) database [3,4] with updated information regarding their substrate specificities, catalytic mechanisms and three-dimensional structures [5]. Each sequence-based family groups together enzymes sharing common structural and mechanistic features. Some families are monospecific with their members acting on only one substrate, but many families group together enzymes with different substrate specificities [3]. Currently, over 165 GH families are listed in the CAZy database and this number keeps on growing [6], reflecting the huge diversity of the enzymatic arsenal devoted to carbohydrate degradation.

Cellulose, the most abundant organic compound on earth, represents an important feedstock for the synthesis of biofuels and platform chemicals [7]. This linear biopolymer composed of β-1,4-linked glucose molecules forms partly crystalline microfibrils, making enzymatic attack difficult [8]. Its enzymatic degradation requires the concerted action of several types of activities [8,9]. Endoglucanases (EC 3.2.1.4) randomly cleave internal β-1,4 glycosidic linkages and increase the number of chain ends; cellobiohydrolases (EC 3.2.1.91) attack the cellulose chain at the reducing or the non-reducing end, processively cleaving every second glycosidic bond to form cellobiose, the smallest structural repeating unit of cellulose; finally, β-glucosidases (EC 3.2.1.21) hydrolyze cellobiose into glucose units. Recently, oxidative cleavage of polysaccharides by Lytic Polysaccharide Monooxygenases (LPMOs) has been demonstrated and shown to enhance the hydrolytic activity of cellulase cocktails [10,11].

GHs and LPMOs often display a modular structure where the catalytic domain (CD) can be appended to one or more carbohydrate-binding modules (CBMs). These auxiliary domains allow binding to insoluble polysaccharides. They are functionally and structurally independent modules, able to target a specific polysaccharide, bringing the CD into close proximity to the substrate [12,13]. CBMs are currently divided into 86 families in the CAZy database. Other accessory modules, such as N-terminal immunoglobulin (Ig)-like domains [14] or fibronectin type III domains [15], can also be associated with CDs. N-terminal Ig-like domains have been shown to interact directly with the CD and to stabilize it in the *Alicyclobacillus acidocaldarius* Cel9A and the *Clostridium thermocellum* cellobiohydrolase CbhA [16,17].

Cellulases are widely distributed in nature and produced by many microorganisms, mainly bacteria and fungi [18], but they are also found in plants [19], some protozoa [20] and some animals [21–23]. Microorganisms have elaborated different strategies for cell wall deconstruction [18]. Aerobic cellulose-degrading bacteria and fungi secrete free cellulases, while anaerobic microbes sometimes express multienzyme complexes termed cellulosomes [8]. In these multienzymatic complexes, a scaffold protein binds various catalytic subunits via cohesion-dockerin interactions [24]. Non-cellulosomal cellulases can also be found in anaerobic microorganisms [25,26], where they are present either in a free form or attached to the cell wall. The advantage of the cellulosomal organization is the proximity of different enzyme activities which generates synergies between enzymes [27]. A similar strategy is applied by the recently characterized thermophilic anaerobic bacterium *Caldicellulosiruptor bescii* which produces multimodular enzymes within a single gene product presenting important synergistic interactions [28].

Cellulolytic bacterial species that have been characterized include the anaerobic *Ruminiclostridium cellulolyticum* and *C. thermocellum* or the aerobic species *Thermobifida fusca* [24,29–33]. Their genomes encode cellulases mainly from families GH5, GH9 and GH48. Many enzymes of the latter family are described as processive GH and bacteria generally contain only a few GH48-encoding genes. In contrast, GH5 and GH9 cellulases which display endoglucanase activities are often more abundant. Thus, the *C. cellulolyticum* genome encodes 13 GH9s and each of the enzymes has a

particular specificity in terms of substrate preference, processivity or binding affinity [34]. GH9 endocellulases often act synergistically with GH48 glucanases, wherein the synergistic factor varies as a function of the GH9 partner enzyme [35,27]. In contrast, the genome of *Clostridium phytofermentans* encodes only one GH9 enzyme that has been shown to be a major contributor to cellulose degradation [36].

In nature, microorganisms often act in a synergistic way to efficiently degrade plant biomass. Previous studies have shown that metagenomes of relevant environments such as soils or compost contain a large number of hemicellulose and cellulose encoding genes, especially if they have been enriched on lignocellulosic substrates [37–39]. Major bacterial phyla present in these aerobic enrichments are *Actinobacteria*, *Bacteroidetes* and *Proteobacteria* [40–42]. These communities thus constitute an interesting reservoir for new enzymes which can be of biotechnological relevance. Concerning the identification of genes involved in lignocellulose degradation, shotgun DNA sequencing delivers a global picture of the functional potential of the microbial community, but suffers from the inconvenience of yielding only very few full-length clones. In contrast, activity-based screening has the advantage of retrieving full coding sequences and identifying genes with low sequence homology to genes of known function [43].

In the present study, a metagenomic library was constructed from compost which had been enriched on lignocellulosic biomass relevant to industrial ethanol production. Screening on model substrates yielded a large number of hemicellulase and cellulose encoding genes, in particular six novel genes encoding GH9 cellulases. The high diversity of still uncharacterized enzymes within the GH9 family, combined with their likely importance in cellulolytic cocktails of aerobic and anaerobic bacteria, led us to characterize and compare three representatives of this family. They originate from two different phyla and are representative of several modular architectures of GH9 enzymes. The results revealed that each of them has distinct biochemical characteristics, including substrate preferences, product profiles and ability to synergize with other enzymes.

2. Materials and methods

2.1 Cloning of metagenome-derived bacterial GH9s for expression in *Escherichia coli*

Compost samples from a municipal compost platform were incubated for seven months with a lignocellulosic substrate consisting of pretreated *Miscanthus giganteus*, wheat straw and poplar and DNA was extracted as previously described [38]. The composition of the lignocellulosic material is detailed in [38]. High molecular weight DNA was separated on a 0.5% low melt agarose gel and the size range of around 40 kb was isolated. After end-repair and a supplementary purification step using Phase Lock Gel (Quantabio, Beverly, MA, USA), the fragments were cloned into pCC2FOS vector and packaged (Epicentre, Madison, WI, USA). The resulting library of about 48,000 clones was screened on AZCL-xyloglucan and AZCL-hydroxycellulose. Clones were grown overnight in microplates with the addition of 0.02 % arabinose to induce a high-copy number of fosmids. After transfer onto LB/chloramphenicol agar plates and overnight growth, an overlay containing lysozyme and AZCL-substrates was applied. Plates were incubated at 55°C for five days to allow enzymatic hydrolysis of the chromogenic substrates [44]. Positive clones were retrieved by observation of clones using a low-power stereo microscope. After a second, identical screening round, positive fosmid clones were sequenced using Illumina sequencing, assembled using the Velvet algorithm and submitted to CAZy family assignment using FASTY (V35) [3,45]. Protein sequences of CAzymes that were relevant for biomass degradation were BLASTed against the non-redundant Protein database for taxonomic assignment using the closest homolog [46]. Full length genes encoding putative GH9 cellulases were subcloned into the pET300/CT-DEST vector using Gateway technology (Thermo Fisher Scientific/Invitrogen, Waltham, MA, USA) and transformed into *E. coli* BL21(DE3) cells (MerckMilliporeSigma, Burlington, MA, USA) for protein production. The full-length gene encoding the unique GH48 identified in the positive fosmid clones was subcloned in a pET22 (MerckMilliporeSigma, Burlington, MA, USA). The primers NdeI-GH48 (5'AAAAAACATATGGCGGTCGCCTGTGACGTGACCTAC3') and XhoI-GH48 (5'AAAAAACTCGAGGGGGAAGAGCAGGCCGTAC3') were designed to amplify the entire gene

by PCR. Using the restriction sites *Nde*I and *Xho*I (underscored in the primers) in a pET22b(+) digested with the same enzymes, the gene was cloned in frame with the sequence encoding a 6xHis-tag at the 3' terminus. The recombinant vector pET-GH48 was used to transform the *E. coli* BL21(DE3) strain for protein production.

2.2 Bioinformatics analysis of the GH9 family

GH9 sequences extracted from the compost metagenome underwent a protein BLAST analysis [46] to determine the taxonomic order of each metagenome-derived GH9 using their respective closest homologs. To construct a phylogenetic tree, the sequences of the 172 characterized GH9s available on the CAZy database on April 17, 2019 were analyzed along with the 6 metagenome-derived GH9s using Geneious software version 9.1.8 (Biomatters, Auckland, New Zealand). Sequences were aligned using Clustal Omega [47] operating with 10 iterations. Only the aligned protein CDs were conserved. Signal peptides and N- and C-terminal extensions that are not fully conserved across the alignment, were removed. Proteins containing partial CDs were excluded from the analysis. The resulting amino acid sequences (168 sequences + 6 metagenome-derived GH9s) were realigned using Clustal Omega. The Neighbor-Joining (NJ) algorithm was used for distance tree building with a bootstrap resampling method using 1000 replicates and a support threshold of 75%. The resulting phylogenetic tree was displayed and annotated using the iTOL online tool [48]. For each sequence, information regarding domain nature and organization, taxonomy and available 3D-structure of the CD was retrieved from the Uniprot [49], Pfam [50] and CDD [51] databases.

2.3 Heterologous expression of the recombinant metagenome-derived GH9s

Recombinant 6xHis-tagged proteins were expressed in *E. coli* BL21 (DE3) cells grown in autoinduction MagicMedia medium (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 100 μg/ml ampicillin. 500 ml cell cultures were grown in 2000 ml flasks at 250 rpm, incubated at 37 °C for 24 h. Cells were harvested by centrifugation at 5000 *g* for 10 min, washed with 0.9% (*w*/*v*) NaCl and stored at -80°C. GH48 enzyme was produced in the recombinant *E. coli* BL21(DE3) strain grown at 37°C to an optical density at 600 nm of 1.5, followed by the addition of 200 μM isopropyl-β-

D-thiogalactopyranoside (IPTG) overnight at 20°C, with shaking. Cells were harvested by centrifuging for15 min at 6000 *g* and processed directly for protein purification.

2.4 Protein purification

All purification steps were carried out at 4°C. Bacterial cell pellets expressing recombinant proteins (GH9b, GH9d and GH9e) were thawed and resuspended in the lysis buffer (25 mM triethanolamine pH 7.0, 150 mM NaCl, 5% (*v*/*v*) glycerol, 15 mM imidazole) supplemented with SigmaFAST Protease Inhibitor Cocktail (1 tablet/50 ml, (MerckMilliporeSigma, Burlington, MA, USA), 1 mg/ml lysozyme (MerckMilliporeSigma, Burlington, MA, USA) and 15 µg/ml DNase I (MerckMilliporeSigma, Burlington, MA, USA). Cells were disrupted by sonication using a Bioblock Scientific Vibra-Cell Ultrasonic Processor (Sonics & Materials Inc). Extracts were spun down at 12 000 *g* for 15 min. Supernatants were loaded onto HisTrap FF crude columns (GE Healthcare, Chicago, IL, USA) pre-equilibrated with the lysis buffer. Washing and elution were carried out in lysis buffer supplemented with imidazole using a two-step gradient (washing at 40 mM imidazole and elution at 112, 88 and 185 mM imidazole for GH9b, GH9d and GH9e, respectively). Purified proteins were concentrated using Vivaspin 20 centrifugal concentrator (30 kDa molecular weight cut-off, Sartorius) and loaded onto a Superdex 200 10/300 GL size-exclusion column pre-equilibrated with gel filtration buffer (25 mM triethanolamine pH 7.0, 150 mM NaCl, 5% (*v*/*v*) glycerol). Elution was carried out using an isocratic flux of buffer. The column was calibrated with Bio-Rad's gel filtration standard (Biorad, Hercules, CA, USA). When size-exclusion chromatography was omitted (GH9e only), imidazole elution buffer was exchanged with gel filtration buffer by diafiltration using the Vivaspin centrifugal concentrator. For GH48 purification, cells were resuspended in 30 mM Tris-HCl (pH 8) and disrupted in a French Press. The crude extract was spun down (10 min, 10000 *g*) and the supernatant containing the His-tagged proteins was loaded onto a column of Ni-nitrilotriacetic acid superflow resin (Qiagen, Hilden, Germany) equilibrated with 30 mM Tris-HCl (pH 8) and eluted using the same buffer supplemented with 60 mM imidazole. The fractions containing the purified protein were pooled and concentrated by ultrafiltration (Vivaspin 10 kDa cutoff, Sartorius, Göttingen, Germany) and loaded onto an anion exchange chromatography column (HiTrap Q-sepharose, GE Healthcare, Chicago, IL, USA)

equilibrated with 30 mM Tris-HCl (pH 8) then eluted with a linear NaCl gradient (0- 0.5 M). Fractions of interest were pooled, dialyzed and concentrated by ultrafiltration (Vivaspin 20, 30 kDa cutoff, Sartorius, Göttingen, Germany) in 30 mM Tris-HCl (pH 8).

Protein concentrations were determined using a Nanodrop spectrophotometer with extinction coefficients calculated using the ExPASy ProtParam tool [52]. Purified proteins were snap-frozen in liquid nitrogen and stored at -80°C for over a year without any significant loss of activity.

2.5 SDS-PAGE

Proteins were separated by SDS-PAGE using 10% polyacrylamide Mini-PROTEAN TGX precast gels (BioRad, Hercules, CA, USA) with Tris-glycine running buffer (BioRad, Hercules, CA, USA) and lithium dodecyl sulfate (LDS) sample buffer (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 50 mM dithiothreitol (DTT). Protein samples were denatured at 70°C for 10 min before loading. The gels were stained with BioSafe Coomassie (Biorad, Hercules, CA, USA).

2.6 Substrates

Carboxymethyl cellulose (CMC) (degree of substitution 0.7) was obtained from SERVA. Low viscosity barley β-glucan, low viscosity Konjac glucomannan and tamarind seed amyloid xyloglucan were purchased from Megazyme. 2-hydroxyethyl-cellulose (HEC) and laminarin were obtained from Merck. Phosphoric acid-swollen cellulose (PASC) was prepared from microcrystalline cellulose Avicel PH-101 (MerckMilliporeSigma, Burlington, MA, USA) as follows: 5 g of Avicel PH-101 was dissolved in 100 ml of 85% phosphoric acid at room temperature and precipitated with cold distilled water. Fibers were collected and washed 6 times in cold distilled water. The pH of the suspension was adjusted to 6.5 with 1 M sodium carbonate. Cellulose was washed again 4 times in cold distilled water. Resulting PASC was collected, drained and stored at - 20°C until use.

2.7 Activity measurements on soluble substrates

Reducing sugar concentration was determined using the 3,5-dinitrosalicylic acid (DNS) assay [53] and a D-glucose standard curve. The effect of pH and temperature on endoglucanase activity was assayed with CMC. Respective enzymes (20 nM to 300 nM) were added to the assay mixture containing 1%

(*w*/*v*) CMC diluted in McIlvaine's citrate-phosphate buffer [54]. The assay mixture was incubated for 10 min. For each enzyme, optimal pH and temperature were respectively assayed in the ranges of 2.7 to 7.4 and 34°C to 85°C. Reactions were stopped by adding 2 volumes of DNS solution (300 g/l potassium sodium tartrate and 10 g/l DNS in 0.4 M sodium hydroxide) and were then heated for 5 min at 95°C. The absorbance was measured at 540 nm using a microplate spectrophotometer (Multiskan ascent, Thermo Fisher Scientific, Waltham, MA, USA).

For stability assays, enzymes were incubated for 24h at different pHs (2.7-7.4) or temperatures (0°C-84°C) and residual activities were measured under optimal conditions for each enzyme.

Specific activities of each enzyme were evaluated on several substrates (low viscosity barley β-glucan, low vicosity Konjac glucomannan, tamarind seed amyloid xyloglucan, 2-hydroxyethyl-cellulose and laminarin) diluted to 1% (*w/v*) in McIlvaine's citrate-phosphate buffer. Respective enzymes (20 nM to 300 nM) were assayed at their corresponding optimal temperatures and pHs.

To determine kinetic parameters on CMC, hydrolysis reactions were carried out at the optimal pH and temperature with a concentration of CMC varying from 0.25% to 4% (*w/v*). The experimental data were fitted to the Michaelis–Menten equation using the maximum likelihood method described in [55]. The Michaelis constant (K_M) and the turn-over number (k_{cat}) were determined from the model. To mesure the effect of different metal ions on the activity, hydrolysis reactions were carried out for 10 min on 1% (*w*/*v*) CMC supplemented or not supplemented with 1 mM $MnCl_2$, $MgCl_2$, $CaCl_2$ or $ZnCl_2$. Respective enzymes (80 nM to 240 nM) were assayed at their corresponding optimal temperatures and pHs.

2.8 Product profile on crystalline and amorphous cellulose

50 nM of each enzyme was incubated at 50°C in the presence of 10 g/l PASC or Avicel PH-101 diluted in McIlvaine's citrate-phosphate buffer, pH 5.9. After 1h, 6h or 24h, the reactions were stopped by additing two volumes of 200 mM NaOH. The mixtures were then filtered at 0.22 µm to remove insoluble substrates. The glucose and corresponding oligosaccharide concentrations were determined using high-performance anion-exchange chromatography (HPAEC) coupled with pulsed

amperometric detection (PAD) (Dionex, Thermo Fisher Scientific, Waltham, MA, USA) and using a CarboPac PA1 column (Thermo Fisher Scientific, Waltham, MA, USA). Data acquisition from the detector and determination of retention times and peak areas were carried out using Chromeleon software version 6.8 (Thermo Fisher Scientific, Waltham, MA, USA). The column was operated at a constant flow rate of 0.3 ml/min at 30 °C. Before injection, the column was equilibrated with 100% of eluent A (0.1 M NaOH) over 5 min. Elution was carried out using a linear gradient of 0–15% of eluent B (1 M sodium acetate in 0.1 M NaOH) over 11 min, followed by a step at 15% of eluent B for 4 min. Glucose and oligosaccharide solutions in the concentration range of 0.1-20 mg/L were used as external standards. Specific activities on Avicel and PASC were inferred from total glucose and oligosaccharide contents.

2.9 Substrate binding assay

Binding of the enzymes to insoluble substrates, Avicel PH-101 and PASC, was assayed using the method carried out by [34]. Briefly, 1.5 µM of each enzyme was incubated at 4°C for 1 h at 400 rpm in 200 µl of cellulose (PASC or Avicel) at 7 g/l in McIlvaine's citrate-phosphate buffer pH 5.9. The samples were subsequently centrifuged at 10 000 *g* for 10 min, and the supernatants were collected and mixed with 3X LDS sample buffer supplemented with 150 mM dithiothreitol. The cellulose-containing pellets were washed twice with McIlvaine's citrate-phosphate buffer, pH 5.9 and resuspended in 200 µl of diluted LDS sample buffer supplemented with 50 mM dithiothreitol. The protein samples were boiled for 10 min and the samples were centrifuged for 10 min at 10 000 *g*. The supernatants were collected and 15 µl was separated by SDS-PAGE as described above.

2.10 Synergy assay

Assays were carried out on Avicel PH-101 (20g/l) with 50 nM of enzyme mixtures at different molar ratios. Avicel PH-101 was hydrolyzed with shaking (850 rpm) for 4h at 50°C and pH 5.9 in McIlvaine's citrate-phosphate buffer. The reactions were stopped by boiling for 10 min. The reducing ends released in the supernatant were assayed by the modified 2,2'-bicinchoninate (BCA) method [56]. Synergistic factors were calculated by dividing the activity of the mixture by the sum of the individual activities.

2.11 Processivity assay

3.5 g/L PASC was hydrolyzed by 1 µM of each enzyme (GH9b, GH9d or GH9e) with shaking (850 rpm) for 6h at 50°C and pH 5.9 in McIlvaine's citrate-phosphate buffer. The reactions were stopped by boiling for 10 min. The supernatant was extracted and the cellulose pellet was washed three times with McIlvaine's buffer. After centrifuging, the reducing ends released in the supernatant and pellet were assayed by the modified BCA method established by [56]. The processivity was determined from the distribution of the reducing ends released in the cellulose pellet (insoluble) and the supernatant (soluble) according to [57].

2.12 Sequence data

All of the sequences in this article were submitted to the GenBank database and have the following accession numbers: MT186814 (GH9a), MT186815 (GH9b), MT186816 (GH9c), MT186817 (GH9d), MT186818 (GH9e), MT186819 (GH9f), MT186820 (GH48).

3. Results

3.1 Sequence and phylogenetic analyses of novel metagenome-derived bacterial GH9 enzymes

Compost samples were enriched on lignocellulosic biomass under aerobic conditions for seven months, after which metagenomic DNA was extracted. A library of 48,000 clones was screened on AZCL-cellulose and AZCL-xyloglucan. After two screening rounds, 74 positive clones were sequenced and ORFs were subjected to CAZy family assignment using the same procedures as those applied for the updates of the CAZy database [4,2]. 31 genes encoding CAZymes potentially involved in plant biomass deconstruction and possessing a signal peptide for secretion could be identified (Supplementary Table S1). Most of them encode glycoside hydrolases, but five carbohydrate esterases from families CE1 and CE4, one AA10 LPMO and two gene fragments containing a carbohydrate binding domain CBM2 were also retrieved. The overall identities with sequences from the non-redundant protein database ranged from 32% (a CE1 carbohydrate esterase) to 92% (for GH74a and

GH51 enzymes). Concerning the probable phylogenetic origin of these genes, two thirds were most closely related to sequences from the *Actinobacteria* phylum, 16% displayed highest identity scores to sequences from the *Bacteroidetes* phylum and two genes were closest to sequences from *Proteobacteria*. Only three genes were related to other phyla. This result is consistent with the observed phylogenetic distribution of the shotgun metagenome from the same enriched compost [38], suggesting that the metagenomic sequences identified here by functional screening might indeed originate from species involved in biomass deconstruction in the present compost community.

For most GH families, only one or two representatives were retrieved from the fosmid library, but notable exceptions are the GH74 and GH9 families, where respectively five and six genes were identified. As GH9 have mostly endo- or exoglucanase activity and are important players in plant biomass deconstruction, these enzymes were investigated further. The six enzymes (denoted GH9a-GH9f) harbored a putative signal peptide suggesting extracellular localization, and they display different modular organizations (Fig. 1). Four of them displayed the highest identity scores (66-83%) with sequences from the *Streptosporangiales* order, whereas GH9b and GH9f showed the most similarity to sequences from a *Thermoflavifilum* species (belonging to the *Chitinophagales* order, 78% identity) and from the genus *Sorangium* (*Myxococcales* order, 69% identity), respectively. All but the GH9b enzyme possess CBMs belonging to families 2, 3 or 4 attached to their catalytic domains. Four proteins harbor a N-terminal Ig-like module located upstream of the CD. GH9a and GH9d display a similar domain architecture as well as a very high sequence identity (94.5% over the full-length sequence and reaching 96% over the CD) (Fig. 1).

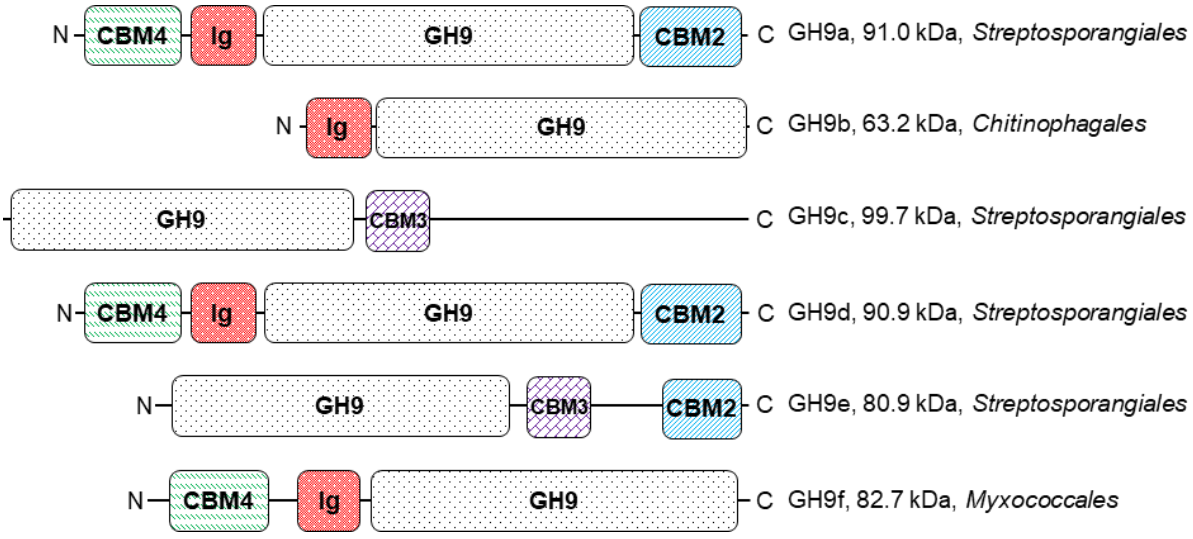


Fig. 1 Domain architecture of the metagenomic-derived bacterial GH9s. The CD (GH9) and accessory modules (N-terminal immunoglobulin-like domain (Ig), CBM2, CBM3 and CBM4) are represented with their relative sequence lengths. The molecular weights of mature proteins were calculated using the ExPASy ProtParam tool [52]. Taxonomic orders were predicted by BLAST [46] using the phylogenetic origin of the closest homolog. N, N-terminal; C, C-terminal

The CDs of our metagenome-derived GH9s were aligned with the 168 characterized GH9s that contain a full-length CD and which were listed in the CAZy database on April 17, 2019. The resulting alignment was used to build a phylogenetic tree (Supplementary Fig. S1 and Fig. 2). Two sub-families appeared upon phylogenetic analysis: GH9 enzymes devoid of an Ig-like module (Pfam accession number PF02927) formed a clade distinct from that containing Ig-like module-containing GH9s. GH9 sequences without an Ig-like module are divided into several taxonomical nodes that separate plant, animal and bacterial proteins with bootstrap values larger than 0.8. Bacterial proteins devoid of an Ig-like module segregate in at least two different clades. Among our metagenome-derived enzymes, Ig-like module-containing GH9a and GH9d, with a 96% sequence identity between their CDs, showed the smallest evolutionary distance. GH9b and GH9f, also harboring an Ig-like module and with, respectively, 48% and 29% sequence identity to GH9d, were clustered in two distinct nodes.

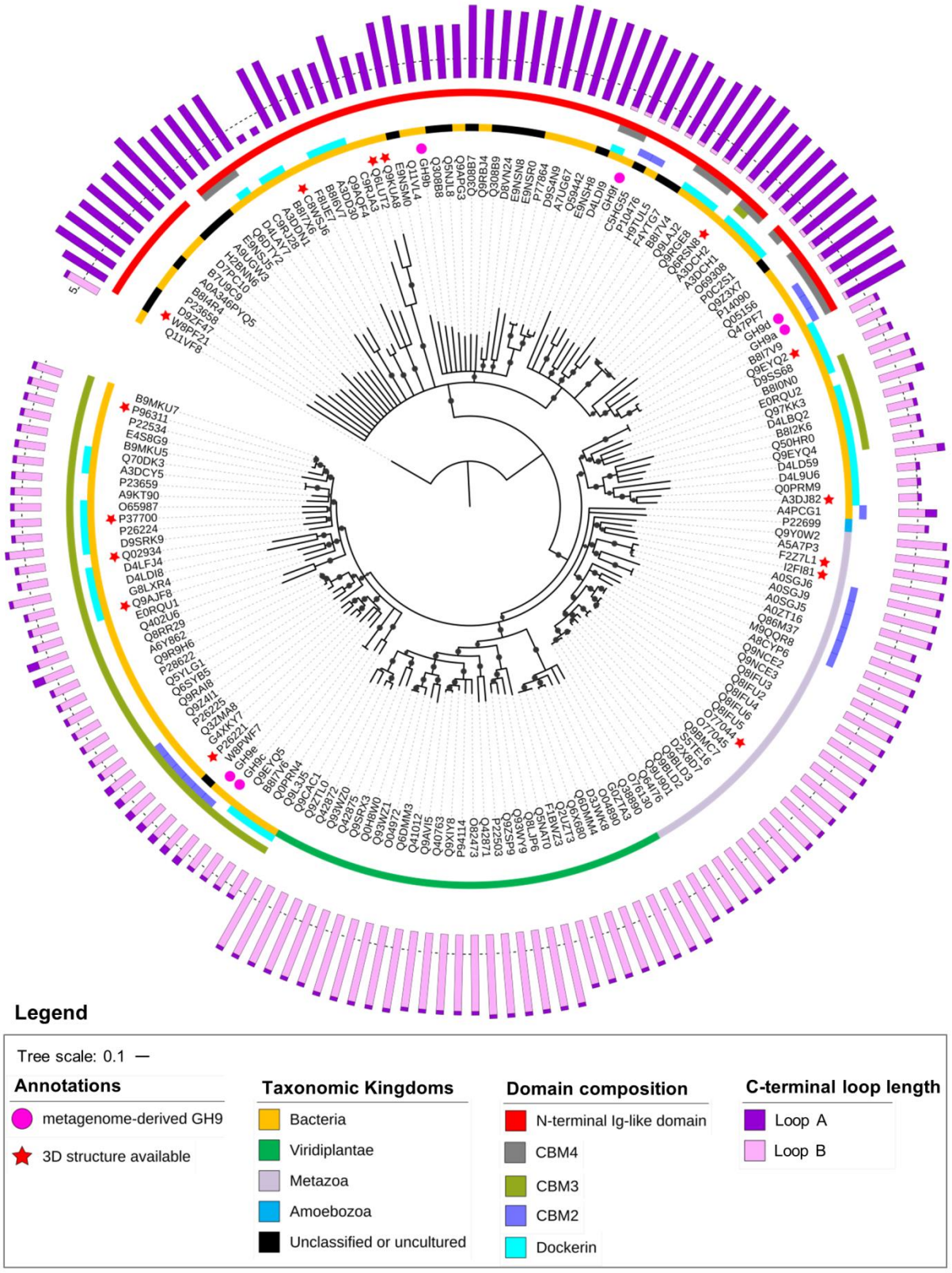

Legend
Tree scale: 0.1
Annotations
metagenome-derived GH9
3D structure available
Taxonomic Kingdoms
Bacteria
Viridiplantae
Metazoa
Amoebozoa
Unclassified or uncultured
Domain composition
N-terminal Ig-like domain
CBM4
CBM3
CBM2
Dockerin
C-terminal loop length
Loop A
Loop B

Fig. 2 Phylogeny of the GH9 family. Circular phylogram based on the alignment of the CDs of the six metagenome-derived GH9s and the 168 characterized GH9 sequences extracted from the CAZy database. The multiple sequence alignment was built using Clustal Omega and the Neighbor-Joining (NJ) algorithm was used for distance tree building. Bootstrap values higher than 80 are represented with a dark dot. Metagenome-derived GH9s are highlighted with a fuchsia dot. Crystallized CDs with a known three-dimensional (3D) structure are highlighted by a red star. The inner ring is colored according to the taxonomic kingdom of the corresponding organism. The middle rings are colored as a function of the presence of one of the five major accessory modules (N-terminal Ig-like domain, CBM2, CBM3, CBM4 or dockerin). The respective lengths of the two possible C-terminal loops are represented in the outer ring as a stacked column chart with a dashed line corresponding to a length of 5 residues. C-terminal loops associated with Ig-like domain-containing GH9s (loop A) or with GH9s exhibiting no Ig-like module (loop B) are respectively represented in violet and in pink. The phylogenetic tree annotation was carried out with iTOL [48]

The tree distance matrix highlighted a low sequence identity between GH9s containing or not containing an Ig-like module (Supplementary Table S2). However, it is likely that they share a common ancestor, as evidenced by residual sequence identity and their similar three-dimensional structures (Supplementary Fig. S2). The CD is a conserved $(\alpha/\alpha)_6$ barrel fold with similar secondary structure orientations. Moreover, our phylogenetic analysis highlighted the presence of a GH9 from *Cytophaga hutchinsonii* (Uniprot accession number Q11VF8), devoid of an Ig-like module and distantly related to both sub-families. Taken together, these results are in line with a common origin for all GH9s and an early loss of the Ig-like domain during evolution.

We could identify 19 different modules in the 174 aligned GH9s highlighting the wide variety of modular architectures in this family (Supplementary Table S3). Among them, only 5 modules are shared by at least 9% of the proteins. All bacterial GH9s, except two, harbored at least one accessory module (i.e. CBM or Ig-like module). CBM4 and CBM9 are closely related and identified under the Pfam accession number PF02018. All of the CBM4/9-containing GH9 enzymes also harbored an Ig-

like domain. In contrast to bacteria, eukaryotic GH9 enzymes, which are all devoid of an Ig-like domain, present much less variability in their modular architecture. Strikingly, they also all lack helper modules, with the exception of some metazoan enzymes comprising a CBM2 module. Two bacterial subgroups can be found among the group of enzymes without Ig-like domain: in one of them, corresponding to the enzymes formerly termed “Theme B” enzymes [58], all enzymes contain a CBM3, which is not true for the second subgroup, where only part of the enzymes contain a CBM3. The second major branch grouping together Ig-like domain-containing enzymes, also displays two subgroups: one of them includes a majority of enzymes with CBM4 (corresponding to the previously termed “Theme D” enzymes). The second one, corresponding to ”Theme C” enzymes, comprises a majority of enzymes without a CBM, but also some enzymes with a CBM4 module. Consequently, the presence or absence of several modules, such as CBM2, CBM3 or dockerin, appeared to be independent of the phylogenetic clustering, suggesting that only the upstream Ig-like module has co-evolved with the CD.

Multiple sequence alignment of the CD highlighted the presence of two distinct C-terminal loops, respectively present in the Ig-like module-containing GH9s (loop A) and in the GH9s from the second branch (loop B) (Fig. 2). Taking a closer look at the structure of two crystallized GH9 members, each belonging to one of the two subfamilies, loop A and B are both bordering the substrate binding cleft (Supplementary Fig. S3). Wu and Davies have already pointed out this difference and explained the different modes of action of the exo-acting GH9 of *Vibrio cholerae* (UniProt: Q9KUA8) and endo-acting GH9s by the presence of a longer loop restricting the active-site binding pocket in the exo-acting enzyme. [59] Our analysis suggests that the longer loop A is present in most Ig-like domain containing enzymes, even if they are endo-acting. Two GH9s without any Ig-like module (UniprotKB accession number A9UGW3 and O69308) appeared among the Ig-like module-containing GH9s and displayed the corresponding loop A. However, these proteins are annotated as partial or fragmentary, according to the CAZy and UniprotKB databases, respectively. Therefore, it is possible that the full-length sequences of these proteins harbor an Ig-like domain.

3.2 Purification of three metagenome-derived bacterial GH9s

In order to characterize the newly identified GH9 enzymes, we intended to subclone them for expression in *E. coli*. For unknown reasons, GH9a and GH9f clones could not be obtained. The strain containing the GH9c-encoding gene did not yield any protein. However, two GH9s harboring an Ig-like domain (GH9b and GH9d) and one GH9 belonging to the branch without an Ig-like domain (GH9e) were successfully cloned and expressed in *E. coli*. Proteins were purified to homogeneity (Supplementary Fig. S4) with estimated purities of 99.5% (GH9b), 91.6% (GH9d), 96.8% (GH9e), as determined by densitometry. Upon separation by size-exclusion chromatography, the proteins displayed apparent molecular weights (MW) that were different from expected (Supplementary Fig. S4). GH9d exhibited the apparent MW of a trimer, assuming it displays a globular shape. GH9b and GH9e MWs appeared respectively 2.4 and 2.1 times smaller, suggesting a non-globular shape. In order to increase yields, GH9e which was poorly expressed but eluted with a high purity rate upon affinity chromatography, was not separated by gel filtration for further experiments.

3.3 Functional characterization on soluble substrates

In order to investigate substrate specificities of metagenome-derived GH9s, specific activities of GH9b, GH9d and GH9e were measured on various soluble substrates (Table 1). The most prominent activities were obtained on β-glucan for the three enzymes. GH9b exhibited the highest activity on all soluble substrates, while GH9e displayed specific activities that were at least 4.2 times lower than for GH9b. All three enzymes showed notable activities on CMC and glucomannan, a linear copolymer of glucose and mannose, joined by β-(1→4)-linkages. In order to evaluate the ability of each enzyme to cleave β-(1→3) linkages, the enzymes were incubated with laminarin, a (1,3)-β-D-glucan with β-(1→6) branching. Compared to β-glucan, a mixed linkage glucose polysaccharide containing β-(1→3) and β-(1→4) bonds, no significant activity was detected on laminarin, indicating that the enzymes are specific for β-(1→4) glycosidic bonds. Overall, specific activities were lower on substrates containing substitutions (CMC, hydroxyethyl cellulose, xyloglucan) compared to non-substituted substrates like β-glucan and glucomannan, suggesting that the functional groups (respectively carboxymethyl, hydroxyethyl and xylose) might cause steric hindrance, thereby rendering access for hydrolases

416 difficult. This might be especially true for xyloglucan due to its bulkier substitution compared to
417 carboxymethyl or hydroxyethyl side chains. Taken together, these results demonstrate that the purified
418 enzymes exhibit endo-β-1,4-glucanase activity.

419

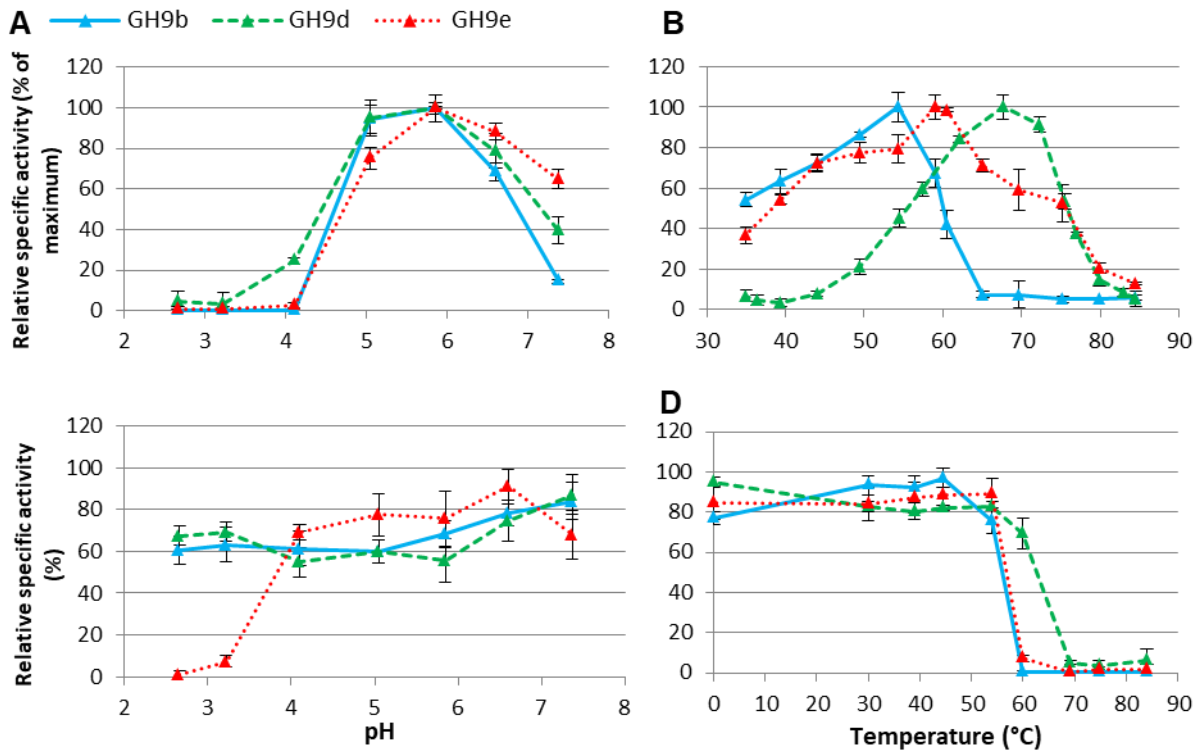
	Specific activities (µmol min^{-1} $nmol^{-1}$)		
	GH9b	GH9d	GH9e
xyloglucan (Tamarind seed)	0.49 ± 0.12	0.15 ± 0.07	0.00 ± 0.00
HEC	0.81 ± 0.01	0.01 ± 0.03	0.09 ± 0.01
CMC	3.94 ± 0.03	1.46 ± 0.07	0.93 ± 0.01
glucomannan (Konjac)	9.94 ± 0.36	3.25 ± 0.47	2.04 ± 0.06
β-glucan (Barley)	26.86 ± 0.74	18.62 ± 0.45	6.13 ± 0.34
Laminarin (*Laminaria digitata*)	0.14 ± 0.12	0.04 ± 0.20	0.03 ± 0.03

420 **Table 1** Specific activities of bacterial GH9s on 1% (*w/v*) soluble substrates. Hydrolysis reactions
421 were carried out on various soluble substrates for 10 min under optimal conditions (pH 5.9 and
422 respective temperatures of 54°C (GH9b), 68°C (GH9d) and 60°C (GH9e) (Fig. 4)). Each value is the
423 mean ± standard deviation of three replicates and is expressed in µmol/min/nmol.

424

425 The effect of pH and temperature on the three endo-β-1,4-glucanases as well as the kinetic parameters
426 were determined on CMC, a model substrate used to characterize cellulolytic activity. A common
427 optimal pH of 5.9 was obtained with rather narrow pH ranges (pH 5 - 6.6) where at least 60% relative
428 activity was observed (Fig. 3A). Regarding pH stability (Fig. 3C), GH9b and GH9d retained at least
429 50% residual activity after 24h of incubation at pHs in the range of 2.7-7.4, whereas GH9e lost its
430 activity below pH 4.1. Variation of the incubation temperature affected the metagenome-derived GH9s
431 differently (Fig. 3B). GH9b displayed an optimal temperature of 54°C, the optimum for GH9e was
432 60°C. GH9d proved to be the most thermophilic enzyme with an optimal temperature of 68°C. GH9e
433 showed more than 50% relative activity over a wide range of temperatures (39-75°C), in contrast to
434 GH9b (35-59°C) and GH9d (58-76°C). Temperature stability profiles (Fig. 3D) highlighted that GH9b

435 and GH9e both lost activity after a 24h incubation above 54°C, while GH9d appeared more stable with
436 a residual activity of 69% at 60°C.



437

438 **Fig. 3** Effect of temperature and pH on the activity of metagenome-derived GH9s on 1% (*w/v*) CMC.
439 (A) Optimal pH. Hydrolysis was carried out at the given pHs and at 68°C (GH9d), 54°C (GH9b) or
440 60°C (GH9e) for 10-20 min. (B) Optimal temperature. Activity was measured after incubation at the
441 indicated temperature and at pH 5.9. (C) pH stability. Activity was measured under optimal conditions
442 (temperature, pH) after a 24h incubation of the enzymes at the designated pH. (D) Temperature
443 stability. Activity was measured under optimal conditions after a 24h incubation at the indicated
444 temperatures. Each value is the mean ± standard deviation of three replicates.

445

446 Kinetic parameters were estimated on CMC (Table 2). Saturation curves fitted the Michaelis-Menten
447 hyperbola. Equation parameters (v_{max} and K_M) were estimated using the maximum likelihood method.
448 Results revealed a lower affinity of GH9e for CMC with a Michaelis constant (K_M) at least 10 times
449 higher than GH9b or GH9d. The turnover numbers (k_{cat}) were close, with a value that was 2 times

450 higher for GH9e. Overall, the catalytic efficiency of GH9e on this soluble substrate is 4.7 to 7 times
451 lower than GH9d and GH9b, respectively.

452

	GH9b	GH9d	GH9e
K_M (g l^{-1})	42.0 ± 2,0	57.1 ± 3.1	574 ± 19
v_{max} (µM min^{-1})	1204 ± 29	310 ± 8.7	7788 ± 177
k_{cat} (s^{-1})	263.9 ± 6.3	241.1 ± 6.8	541.9 ± 12
k_{cat}/K_M (l g^{-1} s^{-1})	6.3 ± 0.5	4.2 ± 0.4	0.9 ± 0.05

453 **Table 2** Kinetic parameters of bacterial GH9s on CMC. Activity was measured under optimal
454 conditions (temperature, pH) for 10 min in a CMC concentration range of 0.25-4% (*w/v*). Each value
455 represents the mean ± standard deviation of two independent experiments, each carried out in
456 triplicate.

457

458 **3.4 Hydrolysis activities on cellulosic substrates**

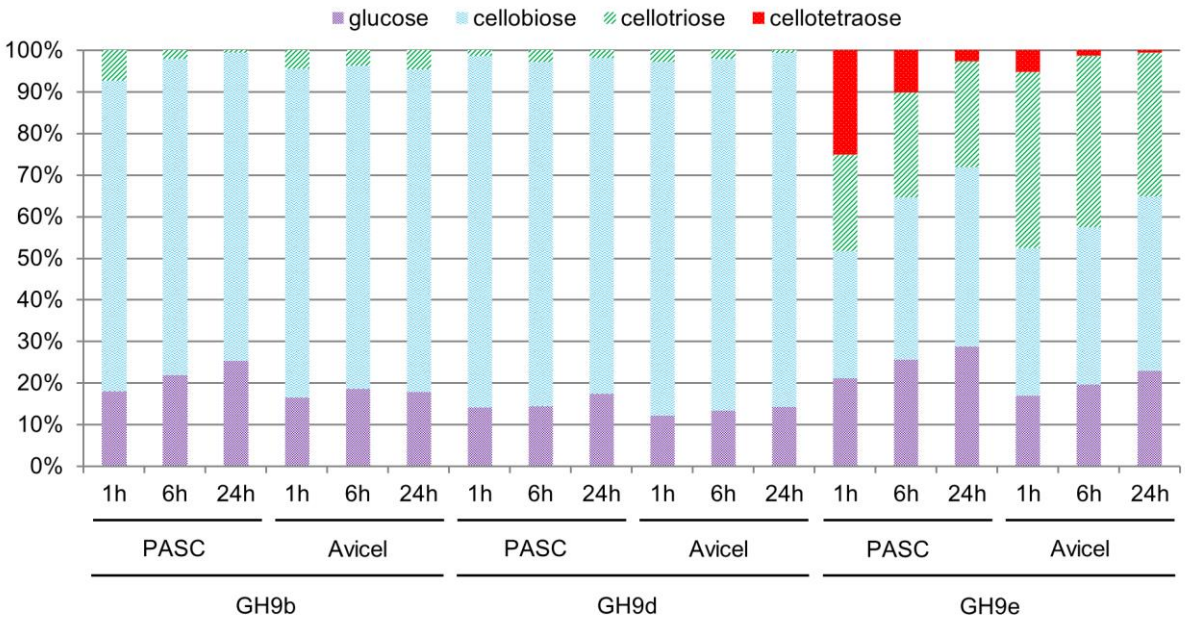
459 We employed lignin-free model biomass substrates, crystalline (Avicel) and amorphous (PASC)
460 cellulose, to compare the product profiles of our metagenome-derived bacterial GH9s. All analyses
461 were carried out at a temperature of 50°C in order to maintain a residual activity of more than 80%
462 over 24h for each enzyme. Results highlighted different propensities to hydrolyze these insoluble
463 substrates (Table 3). GH9e, with the lowest activities measured on soluble substrates, demonstrated
464 the highest activity on Avicel and PASC after 6h and 24 h of incubation. GH9b, with higher specific
465 activities on soluble substrates, exhibited the lowest activities on crystalline cellulose with values 6
466 times lower than GH9e after 24 h of incubation. Such contrasting activities on soluble and insoluble
467 substrates are typical for cellulases and have been already reported [60,61,34]. This discrepancy
468 between the activities of GH9b and GH9e on insoluble substrates (4.4 and 6-fold higher activities of
469 GH9e on PASC and Avicel at 24h, respectively) could be related to the presence of two CBMs in the
470 GH9e enzyme, in contrast to GH9b which does not contain a cellulose binding domain. The GH9d
471 activity appeared only moderate on both insoluble substrates under the present conditions. However,

previous assays on CMC demonstrated that at 50°C, GH9d has only 20% of its maximum activity (Fig. 3B).

Substrate	Incubation time	Hydrolysis activities (µmol $nmol^{-1}$)		
		GH9b	GH9d	GH9e
PASC	1 h	3.38 ± 0.08	0.64 ± 0.04	3.07 ± 0.12
	6 h	3.72 ± 0.08	1.80 ± 0.11	9.15 ± 0.10
	24 h	4.21 ± 0.10	3.09 ± 0.34	18.58 ± 0.35
Avicel	1 h	0.98 ± 0.07	1.05 ± 0.02	2.53 ± 0.07
	6 h	1.40 ± 0.08	1.78 ± 0.03	5.73 ± 0.20
	24 h	1.68 ± 0.09	2.74 ± 0.12	10.38 ± 0.27

Table 3 Hydrolysis activities of bacterial GH9s on 2% (*w/v*) PASC or Avicel are expressed in µmol of reducing sugars liberated per nmol of enzyme after 1 h, 6 h or 24 h at 50°C. The products were quantified by HPAEC-PAD using corresponding external standards. Each value represents the mean ± standard deviation of three replicates.

The product profiles revealed interesting differences (Fig. 4). For GH9b and GH9d, the major product was cellobiose, whereas GH9e displayed a distinct substrate hydrolysis pattern with an elevated proportion of oligosaccharides. Cellotriose and cellotetraose reached 48% of the total sugar content after a one-hour incubation on PASC or Avicel, while cellobiose only reached 31% and 35% of total sugar, respectively. The relative proportions of cellotriose and cellotetraose decreased over time and were respectively reduced to 28% and 35% after a 24h incubation time. These results point to differences in the mode of action of GH9e on the one hand, and GH9b and GH9d on the other hand.



488

489 **Fig. 4** Product profiles on crystalline and amorphous cellulose. 2% (*w/v*) PASC or Avicel were
490 hydrolyzed for 1h, 6h or 24h at pH 5.9 and at 50°C. Products were quantified by HPAEC-PAD using
491 corresponding external standards. The relative content of each compound is shown.

492

493 **3.5 Substrate binding, processivity and synergy on cellulosic substrates**

494 The binding of each GH9 to amorphous and crystalline cellulose was investigated (Fig. 5). No
495 detectable binding of GH9b, which is devoid of CBM, was observed on either substrate. Both CBM-
496 containing GH9s (GH9d and GH9e) bound extensively to amorphous cellulose, especially for GH9d
497 for which no soluble fraction was detectable. The affinity of GH9d and GH9e for crystalline cellulose
498 seemed lower than for amorphous cellulose.

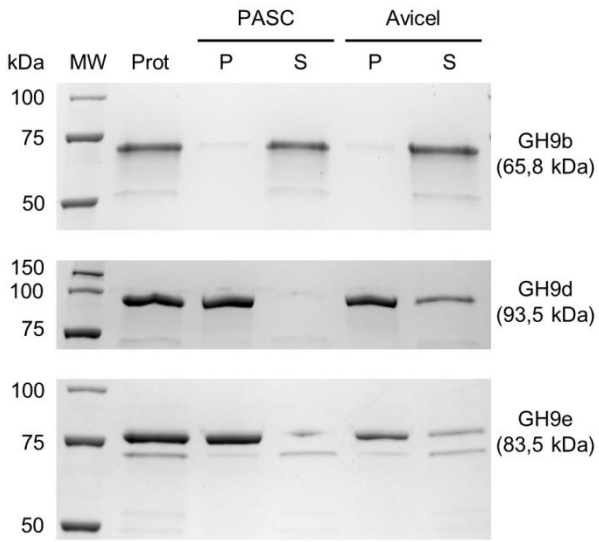


Fig. 5 Binding of GH9 enzymes to cellulosic substrates. Proteins (1.5 µM) were incubated with amorphous (PASC) or crystalline (Avicel) cellulose (7 g/L) for 1 h at 4 °C. The resulting pellets (P, bound proteins) and supernatants (S, unbound protein) were collected and the proteins were denatured in sample buffer. The samples underwent SDS-PAGE along with protein solutions incubated without any substrate (Prot).

We evaluated the processivity of each enzyme using the modified BCA assay, which allows for a quantification of reducing extremities in the micromolar range [56]. We analyzed the distribution of reducing sugars among the insoluble and soluble fractions resulting from the activity of each enzyme on PASC. The vast majority of reducing extremities (>96%) was detected in the soluble fraction of all samples (Table 4). This result suggests that all three enzymes act processively and that a CBM is not necessarily a prerequisite for this mode of action.

Enzyme	Fraction	Reducing ends concentration, µM (%)
GH9b	Soluble	322.6 ± 21.1 (96.4%)
	Insoluble	12.1 ± 23.2 (3.6%)
GH9d	Soluble	222.0 ± 2.4 (97.5%)

	Insoluble	5.7 ± 12.2 (2.5%)
GH9e	Soluble	1554.0 ± 40.0 (99,1%)
	Insoluble	14.1 ± 6.1 (0.9%)

Table 4 Distribution of reducing sugar ends released from PASC. Hydrolysis reactions were carried out with 1 µM of each enzyme for 4 hours at pH 5.9 and 50°C.

The synergistic properties of the three bacterial GH9s were studied between themselves and with a GH48 family member retrieved from the same metagenomic screening (Table 5). Like the GH9 enzymes, this modular protein containing a CBM2 and a GH48 catalytic domain was recombinantly expressed in *E. coli* (Supplementary Fig. S5). Binary and ternary mixtures of the three endoglucanases exhibited moderate synergy between the bacterial GH9s. All enzyme combinations enhanced the activity by a factor of less than 1.5. The highest synergy could be observed by combining GH9b and GH9d. Binary mixtures of GH48 with each of the three GH9s highlighted a significant synergy between GH48 and GH9d. This enzyme combination yielded the highest synergism factor obtained out of all the assayed enzyme mixtures.

Enzyme mixture	Molar ratio in mixture	Synergism factor
GH9b + GH9d	1:1	1.4
GH9b + GH9e	1:1	1.2
GH9d + GH9e	1:1	1.3
GH9b + GH9d + GH9e	1:1:1	1.1
GH9b + GH48	1:1	0.9
GH9d + GH48	1:1	1.6
GH9e + GH48	1:1	0.8

Table 5 Synergy effect of binary or ternary mixtures of bacterial GH9s and in binary mixtures with a GH48 enzyme. Hydrolysis reactions were carried out on Avicel PH-101 for 6 hours at pH 5.9 and 50°C with 50 nM enzymes at different molar ratios. The synergism factor represents the activity of each mixture divided by the sum of the activities of the mixture components.

3.6 Effect of metal ions

It has been shown that cellulase activity can be enhanced or inhibited in the presence of divalent cations [62–65]. We therefore determined the effect of various metal ions (manganese, zinc, magnesium and calcium) on the activity of the three bacterial GH9s (Fig. 6). Endoglucanase activity increased by 70% to 80% in the presence of manganese ions . Calcium ions appeared to be a moderate stimulator of GH9e, with a 10% activity increase, while they had no impact on GH9b and GH9d. Zinc and magnesium ions either slightly inhibited the activity or had no impact.

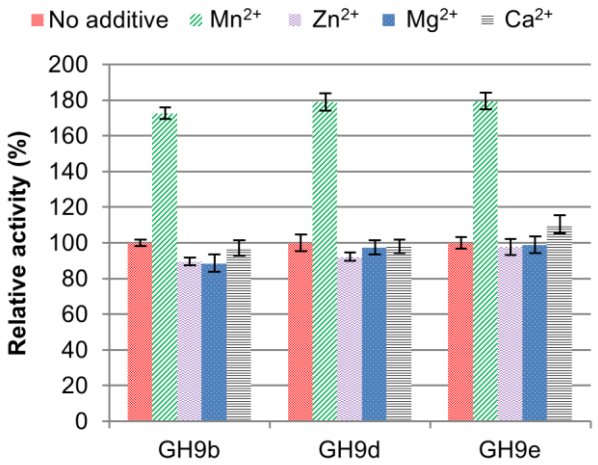


Fig. 6 Effect of 1 mM metal ions on the activity of metagenome-derived GH9s. Hydrolysis reactions were carried out on 1% (w/v) CMC for 10 min under optimal conditions (temperature, pH). Each value is the mean ± standard deviation of three replicates and is expressed as a percentage of the activity of the corresponding enzyme without additive.

4. Discussion

Due to their diversity, environmental bacteria constitute a useful source of promising enzyme candidates for biomass conversion [66]. Compost has been shown to contain a wide diversity of lignocellulose-degrading enzymes which could be enriched by long-term incubation on lignocellulosic biomass [38]. This ecosystem was therefore chosen as a source for the construction of a metagenomic library in order to mine the uncultivable part of cellulose-degrading microflora and to identify novel efficient cellulases. Genes encoding putative cellulases from families GH5, GH6, GH12, GH9 and GH48 were retrieved by functional screening. Compared to other GH families, a surprisingly large number of GH9 encoding genes were identified. This highlights the important role of this class of enzymes in the degradation of lignocellulosic materials by bacteria. GH9 enzymes display various modes of action on cellulosic substrates: some are exoglucanases, such as *C. thermocellum* cellobiohydrolase CbhA [16], while others are processive endocellulases, that remain attached to the polysaccharide chain after the initial hydrolysis step [58,67], or more classical dissociative endoglucanases [68,69]. More rarely, exo-β-D-glucosaminidase, mannanase and xyloglucanase activities have been evidenced for GH9 enzymes [70,71]. In order to understand the role of the GH9 enzymes in our compost enrichment better, a thorough characterization was undertaken.

The sequence comparison of characterized GH9 enzymes carried out in the present study highlighted a wide variety of modular architectures within this enzyme class. In contrast to previous classifications of GH9s [58,72–74], the phylogenetic groups of our analysis, including the newly isolated GH9s, do not reflect the modular architecture of the enzymes. For instance, several domains, like CBM2, CBM3 or dockerin, are each found in several distant clades. The only feature that seems to have co-evolved with the catalytic domain is the N-terminal Ig-like domain, and the two distinct subfamilies are characterized by the absence or presence of this domain, respectively. Among the six metagenomic GH9s identified from our metagenome, three enzymes belonging to different clades were characterized.

GH9e is part of a subgroup of enzymes which all contain a CBM3 module. However, some of them, including GH9e, also possess a CBM2. Of these, two enzymes have previously been extensively characterized: the *Cellulomonas fimi* CenB (UniProt: P26225) [60] and the *Thermobifida fusca* Cel9A (UniProt: P26221) endoglucanases [75,61,76]. The specific activities of GH9e on CMC and PASC are similar to those of both former enzymes [60,61]. All three enzymes seem to be more efficient on PASC than on crystalline cellulose [60,61]. Compared to the other two metagenomic GH9 enzymes characterized in the present study, GH9e is the one with the highest activity on crystalline and amorphous cellulose after a 6h or a 24h hydrolysis.

Within the Ig-like domain containing subgroup which includes GH9d, only very few enzymes possess a CBM2 and a CBM4 module (Fig.2). One of them is GH9d; the other enzymes are GH9a (this study) and UniProt Q47PF7 from *T. fusca* for which no kinetic data have been published. To our knowledge, GH9d thus represents the first GH9 enzyme containing these two modules to be characterized in more detail. *H. thermocellum* CenC is part of the same subgroup as GH9d, but harbors a CBM3 instead of a CBM2. Compared to this enzyme, GH9d displays a higher turnover number on CMC [77]. An interesting feature of GH9d is the similar activity on both PASC and Avicel. This contrasts with the substrate preferences of GH9e and GH9b as well as of other characterized enzymes of their respective subgroups, suggesting that the modular organization of GH9d might contribute to the efficient deconstruction of crystalline cellulose as well as the amorphous substrate by this enzyme.

The representatives of the second subgroup of Ig-like domain-containing enzymes which have been described in most detail are *R. cellulolyticum* Cel9U and Cel9W [34]. GH9b has very similar specific activities to those of the formerly characterized enzymes. GH9b has a higher affinity constant than the *R. cellulolyticum* enzymes, but a comparable $k_{ca}t/K_m$ to Cel9W. The major reaction product of GH9b on Avicel and PASC was cellobiose, suggesting a G2-processive endoglucanase type of action, similar to the closely related *H. thermocellum* Cel9D [78].

GH9 enzymes have previously also been isolated from other metagenomes, such as soil or compost metagenomes [62,79]. Most of them are Ig-like domain containing enzymes devoid of a CBM.

Similarly to GH9b, they displayed high activity on soluble substrates, such as CMC or β-glucan [80,62,79]. The Cel9 endoglucanase from a bagasse pile metagenome with the same modular structure as GH9e showed high activity on β-glucan, but not on the other cellulosic substrates tested [81].

All three GH9 enzymes had optimal temperatures above 50°C and a pronounced thermostability below 60°C, with GH9d being the most thermophilic and thermostable endoglucanase. These results are in line with activities measured for secreted or cellulosomal GH9 family counterparts presenting heterogeneous modular organizations [77,82,83,63]. The reported optimal temperatures ranged from 40°C for EngZ from *C. cellulovorans* to 70°C for *H. thermocellum* CenC, with thermostability up to at least 50°C [83,77]. Optimal temperatures for previously identified metagenomic GH9 enzymes were shown to differ from 30°C to 75°C, although the enzymes have been retrieved from mesophilic environments [81,62,80].

All of the three GH9 enzymes displayed Michaelis-Menten kinetics on CMC, despite the processive activity observed. CMC is a substituted substrate that might hinder any processive action, as demonstrated by the absence of activity of cellobiohydrolase on this soluble substrate. Therefore, CMC could force endoglucanases to detach from the substrate in order to bind and liberate a new product. This condition might be necessary in order to obey the Michaelis-Menten model.

The GH9e enzyme displayed a distinct product profile on cellulose with a high relative content of cellotetraose (G4) and cellotriose (G3) oligosaccharides. A similar cellulose hydrolysis pattern was previously obtained with *T. fusca* Cel9A [84,85]. Our phylogenetic analysis highlighted that GH9e CD is closely related to *T. fusca* Cel9A (76% identity). Both proteins show a similar modular organization and an ability to processively hydrolyze cellulosic substrates and produce oligosaccharides. Cellulases with this particular hydrolysis pattern are found in both cellulosomal [78,86] and non-cellulosomal GH9s [87,84,88]. In their comparative study of the cellulosomal cellulases from *C. thermocellum*, Leis *et al*. [78] identified G4-type processive endoglucanases only within the GH9 family. These enzymes

catalyze the cleavage and release of defined cello-oligosaccharides with G4 as an intermediate product at the beginning of the hydrolysis reaction.

In our phylogenetic tree, characterized G4-producing GH9s with a processive activity [87,84,88,78,58,67] clustered in the same clade as GH9e, which also displayed processivity. These G4-producing GH9s and the metagenome-derived GH9c and GH9e harbor a CBM3 found C-terminal to the CD. In the vast majority of G4-producing GH9s, including GH9e, this CBM3 belongs to type c (CBM3c) based on sequence identity [74]. CBM3cs do not exhibit all of the conserved residues found in type a and b CBM3s, including residues involved in cellulose-binding [58]. Furthermore, a truncated form of *H. thermocellum* Cel9I composed of only the catalytic domain and a CBM3c module failed to bind to cellulose [58]. Rather, this CBM appears to play a helper role in activity, as cellulose degradation is markedly reduced and processivity is lost in *T. fusca* Cel9A upon deletion of its CBM3c [85]. CBM3cs are therefore thought to be involved in modulating the activity and processivity of processive GH9 endoglucanases.

The role of other CBMs on processivity is less clear and might depend on other structural features of the enzymes. The processivity of *T. fusca* Cel9A was reduced upon deletion of its CBM2 and a role of the CBMX2s of *C. cellulosi* Cel9a in processivity has also been evidenced [85,89]. The mechanisms involved in processivity for GH9 enzymes without CBM3c are still less well understood [90]. In the present study, the processive mode of action observed on amorphous cellulose for all three GH9 enzymes, including GH9b devoid of CBM, suggests that the presence of a CBM is not a necessary prerequisite for processivity. However, deletion of CBM2 and/or CBM4 should be carried out to determine the role of these CBMs in processivity for GH9d and GH9e.

Manganese ion is an activator of all three bacterial GH9s assayed on CMC. As the CD is the only common module of our three bacterial endoglucanases, this result suggests a potential stabilizing role for manganese ions on the CDs of GH9 enzymes. However, supplementing with manganese has already been assayed on CMC for several GH9 family members, with contrasting results. Mn^{2+} significantly enhanced the activity of a termite GH9 [64] and of another GH9 cloned from a compost

metagenome [62]. It had no positive impact on *C. thermocellum* Cel9W activity [65] while it significantly inhibited the activity of Cel9K from *Paenibacillus* sp. X4 [63]. Altogether, this suggests that discrete factors, not fully conserved, might influence manganese binding. Further experiments such as differential scanning fluorimetry would be necessary in order to detect ligand interactions that might promote protein stability [91].

Of our three GH9s, the highest synergy on Avicel cellulose was obtained between GH9b and GH9d. Each of these enzymes had limited activity on crystalline cellulose, but they displayed contrasting binding abilities. It is possible that, due to the presence of both a CBM2 and a CBM4, GH9d binds to specific sites of the substrates and renders it more accessible for GH9b. GH9e, which harbors a CBM2 and a CBM3 at its C-terminus, might not be as efficient as GH9d in creating accessible sites for GH9b, thus leading to lower synergy with GH9b. Alternatively, the high activity of GH9e compared with the other two enzymes (about 2.5 to 6 times higher, depending on hydrolysis time) might explain the more limited synergistic properties observed with GH9e. Similar synergy factors than with GH9b were found for GH9d and GH9e in combination with the GH48 family member retrieved from our compost metagenome. The best synergy was observed by mixing GH9d and GH48, which is in good agreement with the previously demonstrated synergy between GH9 and GH48 enzymes [35,27].

5. Conclusions

A wide diversity of GH9 enzymes was retrieved in a compost metagenome and three new bacterial members of this family were characterized as thermophilic and processive endo-1,4-β-glucanases activated by manganese. The highest activity on insoluble substrates was achieved by GH9d and GH9e containing a CBM. GH9e displayed a peculiar processive activity, rapidly producing cellotetraose and cellotriose from amorphous cellulose. GH9d, one of the first GH9 enzymes with CBM4 and a CBM2 to be characterized, has various interesting features: high stability in a wide range of pH and temperatures, similar activity on soluble and insoluble substrates, and the highest synergy

with GH9b and GH48 enzymes. These properties could make it suitable for biomass saccharification for bioethanol production. From a more fundamental point of view, it will be interesting to determine the relative contribution of the two binding modules to affinity, activity and processivity on diverse substrates by means of deletion mutants and to characterize more representatives of the same modular structure in further studies in order to reveal any common features.

Acknowledgments

The authors thank Nicolas Lopes Ferreira for fruitful discussions and for providing PASC, as well as Simon Arragain for critical review of the manuscript. This work was supported by Agence Nationale de Recherche (Grant number: ANR-12-BIO-ME-006-01)

Authors' contributions

LA and NF purified and characterized the enzymes, and LA drafted the manuscript. AH constructed the metagenomic library and isolated positive clones. BH and VL carried out the CAZyme annotation and BH proofread the manuscript. EJ and LA carried out the phylogenetic analysis. SHB designed the study and SP and SHB supervised the work and participated in manuscript drafting and editing.

6. References

[1]N. Karaki, A. Aljawish, C. Humeau, L. Muniglia, J. Jasniewski, Enzymatic modification of polysaccharides: Mechanisms, properties, and potential applications: A review, Enzyme and Microbial Technology 90 (2016) 1–18.

[2]B.L. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): An expert resource for Glycogenomics, Nucleic acids research 37 (2009) D233-D238.

[3]B. Henrissat, G. Davies, Structural and sequence-based classification of glycoside hydrolases, Current opinion in structural biology 7 (1997) 637–644.

[4]V. Lombard, H. Golaconda Ramulu, E. Drula, P.M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic acids research 42 (2014) D490-5.

[5]G. Davies, B. Henrissat, Structures and mechanisms of glycosyl hydrolases, Structure (London, England : 1993) 3 (1995) 853–859.

[6]M.-L. Garron, B. Henrissat, The continuing expansion of CAZymes and their families, Current opinion in chemical biology 53 (2019) 82–87.

[7]L.R. Lynd, C.E. Wyman, T.U. Gerngross, Biocommodity Engineering, Biotechnol Progress 15 (1999) 777–793.

[8]L.R. Lynd, P.J. Weimer, W.H. van Zyl, I.S. Pretorius, Microbial cellulose utilization: Fundamentals and biotechnology, Microbiology and molecular biology reviews : MMBR 66 (2002) 506-77, table of contents.

[9]Y.-H.P. Zhang, L.R. Lynd, Toward an aggregated understanding of enzymatic hydrolysis of cellulose: Noncomplexed cellulase systems, Biotechnology and bioengineering 88 (2004) 797–824.

[10]P.V. Harris, D. Welner, K.C. McFarland, E. Re, J.-C. Navarro Poulsen, K. Brown, R. Salbo, H. Ding, E. Vlasenko, S. Merino, F. Xu, J. Cherry, S. Larsen, L. Lo Leggio, Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: Structure and function of a large, enigmatic family, Biochemistry 49 (2010) 3305–3316.

[11]S.J. Horn, G. Vaaje-Kolstad, B. Westereng, V.G. Eijsink, Novel enzymes for the degradation of cellulose, Biotechnology for biofuels 5 (2012) 45.

[12]A.B. Boraston, D.N. Bolam, H.J. Gilbert, G.J. Davies, Carbohydrate-binding modules: Fine-tuning polysaccharide recognition, The Biochemical journal 382 (2004) 769–781.

[13]D. Guillén, S. Sánchez, R. Rodríguez-Sanoja, Carbohydrate-binding domains: Multiplicity of biological roles, Applied microbiology and biotechnology 85 (2010) 1241–1249.

[14]M. Juy, A.G. Amrt, P.M. Alzari, R.J. Poljak, M. Claeyssens, P. Béguin, J.-P. Aubert, Three-dimensional structure of a thermostable bacterial cellulase, Nature 357 (1992) 89–91.

[15]I.A. Kataeva, R.D. Seidel, A. Shah, L.T. West, X.-L. Li, L.G. Ljungdahl, The fibronectin type 3-like repeat from the Clostridium thermocellum cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface, Applied and environmental microbiology 68 (2002) 4292–4300.

[16]I.A. Kataeva, V.N. Uversky, J.M. Brewer, F. Schubot, J.P. Rose, B.-C. Wang, L.G. Ljungdahl, Interactions between immunoglobulin-like and catalytic modules in Clostridium thermocellum cellulosomal cellobiohydrolase CbhA, Protein engineering, design & selection : PEDS 17 (2004) 759–769.

[17]H. Liu, J.H. Pereira, P.D. Adams, R. Sapra, B.A. Simmons, K.L. Sale, Molecular simulations provide new insights into the role of the accessory immunoglobulin-like domain of Cel9A, FEBS letters 584 (2010) 3431–3435.

[18]Y.J. Bomble, C.-Y. Lin, A. Amore, H. Wei, E.K. Holwerda, P.N. Ciesielski, B.S. Donohoe, S.R. Decker, L.R. Lynd, M.E. Himmel, Lignocellulose deconstruction in the biosphere, Current opinion in chemical biology 41 (2017) 61–70.

[19]B.R. Urbanowicz, A.B. Bennett, E. Del Campillo, C. Catalá, T. Hayashi, B. Henrissat, H. Höfte, S.J. McQueen-Mason, S.E. Patterson, O. Shoseyov, T.T. Teeri, J.K.C. Rose, Structural organization and a standardized nomenclature for plant endo-1,4-beta-glucanases (cellulases) of glycosyl hydrolase family 9, Plant physiology 144 (2007) 1693–1696.

[20]R. Ramalingam, J.E. Blume, H.L. Ennis, The Dictyostelium discoideum spore germination-specific cellulase is organized into functional domains, Journal of bacteriology 174 (1992) 7834–7837.

[21]K. Suzuki, T. Ojima, K. Nishita, Purification and cDNA cloning of a cellulase from abalone Haliotis discus hannai, European journal of biochemistry 270 (2003) 771–778.

[22]T. Arimori, A. Ito, M. Nakazawa, M. Ueda, T. Tamada, Crystal structure of endo-1,4-β-glucanase from Eisenia fetida, Journal of synchrotron radiation 20 (2013) 884–889.

[23]S. Khademi, L.A. Guarino, H. Watanabe, G. Tokuda, E.F. Meyer, Structure of an endoglucanase from termite, Nasutitermes takasagoensis, Acta crystallographica. Section D, Biological crystallography 58 (2002) 653–659.

[24]W. Schwarz, The cellulosome and cellulose degradation by anaerobic bacteria, Applied microbiology and biotechnology 56 (2001) 634–649.

[25]E. Berger, D. Zhang, V.V. Zverlov, W.H. Schwarz, Two noncellulosomal cellulases of Clostridium thermocellum, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically, FEMS microbiology letters 268 (2007) 194–201.

[26]N. Franche, C. Tardif, J. Ravachol, S. Harchouni, P.-H. Ferdinand, R. Borne, H.-P. Fierobe, S. Perret, Cel5I, a SLH-Containing Glycoside Hydrolase: Characterization and Investigation on Its Role in Ruminiclostridium cellulolyticum, Plos One 11 (2016) e0160812.

[27]H.-P. Fierobe, E.A. Bayer, C. Tardif, M. Czjzek, A. Mechaly, A. Bélaïch, R. Lamed, Y. Shoham, J.-P. Bélaïch, Degradation of Cellulose Substrates by Cellulosome Chimeras: SUBSTRATE TARGETING VERSUS PROXIMITY OF ENZYME COMPONENTS, Journal of Biological Chemistry 277 (2002) 49621–49630.

[28]R. Brunecky, M. Alahuhta, Q. Xu, B.S. Donohoe, M.F. Crowley, I.A. Kataeva, S.-J. Yang, M.G. Resch, M.W.W. Adams, V.V. Lunin, M.E. Himmel, Y.J. Bomble, Revealing nature's cellulase diversity: The digestion mechanism of Caldicellulosiruptor bescii CelA, Science (New York, N.Y.) 342 (2013) 1513–1516.

[29]M.K. Bhat, T.M. Wood, The cellulase of the anaerobic bacterium Clostridium thermocellum: Isolation, dissociation, and reassociation of the cellulosome, Carbohydrate Research 227 (1992) 293–300.

768 [30]I. Fendri, C. Tardif, H.-P. Fierobe, S. Lignon, O. Valette, S. Pagès, S. Perret, The cellulosomes
769 from Clostridium cellulolyticum : Identification of new components and synergies between
770 complexes, The FEBS Journal 276 (2009) 3076–3086.

771 [31]E.M. Gomez del Pulgar, A. Saadeddin, The cellulolytic system of Thermobifida fusca, Critical
772 Reviews in Microbiology 40 (2014) 236–247.

773 [32]J.-C. Blouzard, P.M. Coutinho, H.-P. Fierobe, B. Henrissat, S. Lignon, C. Tardif, S. Pagès, P. de
774 Philip, Modulation of cellulosome composition in Clostridium cellulolyticum: Adaptation to the
775 polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses,
776 Proteomics 10 (2010) 541–554.

777 [33]J.L.A. Brás, A. Cartmell, A.L.M. Carvalho, G. Verzé, E.A. Bayer, Y. Vazana, M.A.S. Correia,
778 J.A.M. Prates, S. Ratnaparkhe, A.B. Boraston, M.J. Romão, C.M.G.A. Fontes, H.J. Gilbert, Structural
779 insights into a unique cellulase fold and mechanism of cellulose hydrolysis, Proceedings of the
780 National Academy of Sciences of the United States of America 108 (2011) 5237–5242.

781 [34]J. Ravachol, R. Borne, C. Tardif, P. de Philip, H.-P. Fierobe, Characterization of all family-9
782 glycoside hydrolases synthesized by the cellulosome-producing bacterium Clostridium cellulolyticum,
783 The Journal of biological chemistry 289 (2014) 7335–7348.

784 [35]M. Kostylev, D.B. Wilson, Synergistic interactions in cellulos hydrolysis, Biofuels 3 (2012) 61–
785 70.

786 [36]A.C. Tolonen, A.C. Chilaka, G.M. Church, Targeted gene inactivation in Clostridium
787 phytofermentans shows that cellulose degradation requires the family 9 hydrolase Cphy3367,
788 Molecular microbiology 74 (2009) 1300–1313.

789 [37]T. Mori, I. Kamei, H. Hirai, R. Kondo, Identification of novel glycosyl hydrolases with
790 cellulolytic activity against crystalline cellulose from metagenomic libraries constructed from bacterial
791 enrichment cultures, Springerplus 3 (2014) 365.

792 [38]S. Heiss-Blanquet, F. Fayolle-Guichard, V. Lombard, A. Hébert, P.M. Coutinho, A. Groppi, A.
793 Barre, B. Henrissat, Composting-Like Conditions Are More Efficient for Enrichment and Diversity of

Organisms Containing Cellulase-Encoding Genes than Submerged Cultures, Plos One 11 (2016) e0167216-e0167216.

[39]K.M. DeAngelis, J.M. Gladden, M. Allgaier, P. D'haeseleer, J.L. Fortney, A. Reddy, P. Hugenholtz, S.W. Singer, J.S. Vander Gheynst, W.L. Silver, B.A. Simmons, T.C. Hazen, Strategies for Enhancing the Effectiveness of Metagenomic-based Enzyme Discovery in Lignocellulolytic Microbial Communities, BioEnergy Research 3 (2010) 146–158.

[40]W. Mhuantong, V. Charoensawan, P. Kanokratana, S. Tangphatsornruang, V. Champreda, Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass-degrading enzymes among lignocellulolytic microbial communities, Biotechnology for biofuels 8 (2015) 16.

[41]M. de Vries, A. Schöler, J. Ertl, Z. Xu, M. Schloter, Metagenomic analyses reveal no differences in genes involved in cellulose degradation under different tillage treatments, FEMS microbiology ecology 91 (2015).

[42]A.P. Reddy, C.W. Simmons, P. D'haeseleer, J. Khudyakov, H. Burd, M. Hadi, B.A. Simmons, S.W. Singer, M.P. Thelen, J.S. Vandergheynst, Discovery of microorganisms and enzymes involved in high-solids decomposition of rice straw using metagenomic analyses, Plos One 8 (2013) e77985-e77985.

[43]C. Simon, R. Daniel, Achievements and new knowledge unraveled by metagenomic approaches, Applied microbiology and biotechnology 85 (2009) 265–276.

[44]M. Nyyssönen, H.M. Tran, U. Karaoz, C. Weihe, M.Z. Hadi, J.B.H. Martiny, A.C. Martiny, E.L. Brodie, Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries, Front Microbiol 4 (2013) 282.

[45]W.R. Pearson, T. Wood, Z. Zhang, W. Miller, Comparison of DNA Sequences with Protein Sequences, Genomics 46 (1997) 24–36.

[46]S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic acids research 25 (1997) 3389–3402.

[47]F. Madeira, Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A.R.N. Tivey, S.C. Potter, R.D. Finn, R. Lopez, The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic acids research 47 (2019) W636-W641.

[48]I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: Recent updates and new developments, Nucleic acids research 47 (2019) W256-W259.

[49]UniProt Consortium, UniProt: A worldwide hub of protein knowledge, Nucleic acids research 47 (2019) D506-D515.

[50]R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, A. Bateman, Pfam: Clans, web tools and services, Nucleic acids research 34 (2006) D247-51.

[51]A. Marchler-Bauer, Y. Bo, L. Han, J. He, C.J. Lanczycki, S. Lu, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng, L.Y. Geer, S.H. Bryant, CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures, Nucleic acids research 45 (2017) D200-D203.

[52]E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, Protein Identification and Analysis Tools on the ExPASy Server, in: J.M. Walker (Ed.), The Proteomics Protocols Handbook, Humana Press; Springer e-books, Totowa, NJ, 2005.

[53]G.L. Miller, Use of Dinitrosalicylic Acid Reagent for Determination of Reducing Sugar, Anal. Chem. 31 (1959) 426–428.

[54]T.C. McIlvaine, A buffer solution for colorimetric comparison, J. Biol. Chem. 49 (1921) 183–186.

[55]J.G. Raaijmakers, Statistical analysis of the Michaelis-Menten equation, Biometrics 43 (1987) 793–803.

[56]Y.-H.P. Zhang, L.R. Lynd, Determination of the number-average degree of polymerization of cellodextrins and cellulose with application to enzymatic hydrolysis, Biomacromolecules 6 (2005) 1510–1515.

[57]D.C. Irwin, M. Spezio, L.P. Walker, D.B. Wilson, Activity studies of eight purified cellulases: Specificity, synergism, and binding domain effects, Biotechnology and bioengineering 42 (1993) 1002–1013.

[58]R. Gilad, L. Rabinovich, S. Yaron, E.A. Bayer, R. Lamed, H.J. Gilbert, Y. Shoham, CelI, a noncellulosomal family 9 enzyme from Clostridium thermocellum, is a processive endoglucanase that degrades crystalline cellulose, Journal of bacteriology 185 (2003) 391–398.

[59]L. Wu, G.J. Davies, Structure of the GH9 glucosidase/glucosaminidase from Vibrio cholerae, Acta Crystallogr F Struct Biol Commun 74 (2018) 512–523.

[60]P. Tomme, A.L. Creagh, D.G. Kilburn, C.A. Haynes, Interaction of Polysaccharides with the N-Terminal Cellulose-Binding Domain of Cellulomonas fimi CenC. 1. Binding Specificity and Calorimetric Analysis, Biochemistry 35 (1996) 13885–13894.

[61]W. Zhou, D.C. Irwin, J. Escovar-Kousen, D.B. Wilson, Kinetic Studies of Thermobifida fusca Cel9A Active Site Mutant Enzymes, Biochemistry 43 (2004) 9655–9663.

[62]C.-J. Duan, M.-Y. Huang, H. Pang, J. Zhao, C.-X. Wu, J.-X. Feng, Characterization of a novel theme C glycoside hydrolase family 9 cellulase and its CBM-chimeric enzymes, Applied microbiology and biotechnology 101 (2017) 5723–5737.

[63]J.P. Lee, Y.A. Kim, S.K. Kim, H. Kim, Characterization of a Multimodular Endo-β-1,4-Glucanase (Cel9K) from Paenibacillus sp. X4 with a Potential Additive for Saccharification, Journal of microbiology and biotechnology 28 (2018) 588–596.

[64]P. Zhang, X. Yuan, Y. Du, J.-J. Li, Heterologous expression and biochemical characterization of a GHF9 endoglucanase from the termite Reticulitermes speratus in Pichia pastoris, BMC biotechnology 18 (2018) 35.

[65]K. Kumar, S. Singal, A. Goyal, Role of carbohydrate binding module (CBM3c) of GH9 β-1,4 endoglucanase (Cel9W) from Hungateiclostridium thermocellum ATCC 27405 in catalysis, Carbohydrate Research 484 (2019) 107782.

[66]R. López-Mondéjar, C. Algora, P. Baldrian, Lignocellulolytic systems of soil bacteria: A vast and diverse toolbox for biotechnological conversion processes, Biotechnology advances (2019).

[67]X.-Z. Zhang, N. Sathitsuksanoh, Y.-H.P. Zhang, Glycoside hydrolase family 9 processive endoglucanase from Clostridium phytofermentans: Heterologous expression, characterization, and synergy with family 48 cellobiohydrolase, Bioresource technology 101 (2010) 5534–5538.

[68]M.M. Kesavulu, J.Y. Tsai, H.L. Lee, P.H. Liang, C.D. Hsiao, Structure of the catalytic domain of the Clostridium thermocellum cellulase CelT, Acta crystallographica. Section D, Biological crystallography 68 (2012) 310–320.

[69]G. Parsiegla, A. Belaïch, J.P. Belaïch, R. Haser, Crystal structure of the cellulase Cel9M enlightens structure/function relationships of the variable catalytic modules in glycoside hydrolases, Biochemistry 41 (2002) 11134–11142.

[70]Y. Honda, N. Shimaya, K. Ishisaki, M. Ebihara, H. Taniguchi, Elucidation of exo-beta-D-glucosaminidase activity of a family 9 glycoside hydrolase (PBPRA0520) from Photobacterium profundum SS9, Glycobiology 21 (2011) 503–511.

[71]V. Phakeenuya, K. Ratanakhanokchai, A. Kosugi, C. Tachaapaikoon, A novel multifunctional GH9 enzyme from Paenibacillus curdlanolyticus B-6 exhibiting endo/exo functions of cellulase, mannanase and xylanase activities, Applied microbiology and biotechnology 104 (2020) 2079–2096.

[72]A. Belaich, G. Parsiegla, L. Gal, C. Villard, R. Haser, J.-P. Belaich, Cel9M, a new family 9 cellulase of the Clostridium cellulolyticum cellulosome, Journal of bacteriology 184 (2002) 1378–1384.

[73]S.Y. Ding, E.A. Bayer, D. Steiner, Y. Shoham, R. Lamed, A novel cellulosomal scaffoldin from Acetivibrio cellulolyticus that contains a family 9 glycosyl hydrolase, Journal of bacteriology 181 (1999) 6720–6729.

[74]S. Jindou, Q. Xu, R. Kenig, M. Shulman, Y. Shoham, E.A. Bayer, R. Lamed, Novel architecture of family-9 glycoside hydrolases identified in cellulosomal enzymes of Acetivibrio cellulolyticus and Clostridium thermocellum, FEMS microbiology letters 254 (2006) 308–316.

[75]F. Mingardon, J.D. Bagert, C. Maisonnier, D.L. Trudeau, F.H. Arnold, Comparison of family 9 cellulases from mesophilic and thermophilic bacteria, Appl Environ Microbiol 77 (2011) 1436–1442.

[76]D.B. Wilson, Studies of Thermobifida fusca plant cell wall degrading enzymes, Chem Record 4 (2004) 72–82.

[77]I.u. Haq, F. Akram, M.A. Khan, Z. Hussain, A. Nawaz, K. Iqbal, A.J. Shah, CenC, a multidomain thermostable GH9 processive endoglucanase from Clostridium thermocellum: Cloning, characterization and saccharification studies, World journal of microbiology & biotechnology 31 (2015) 1699–1710.

[78]B. Leis, C. Held, F. Bergkemper, K. Dennemarck, R. Steinbauer, A. Reiter, M. Mechelke, M. Moerch, S. Graubner, W. Liebl, W.H. Schwarz, V.V. Zverlov, Comparative characterization of all cellulosomal cellulases from Clostridium thermocellum reveals high diversity in endoglucanase product formation essential for complex activity, Biotechnology for biofuels 10 (2017) 240.

[79]Y. Zhou, X. Wang, W. Wei, J. Xu, W. Wang, Z. Xie, Z. Zhang, H. Jiang, Q. Wang, C. Wei, A novel efficient β-glucanase from a paddy soil microbial metagenome with versatile activities, Biotechnology for biofuels 9 (2016) 36.

[80]H. Nacke, M. Engelhaupt, S. Brady, C. Fischer, J. Tautzt, R. Daniel, Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes, Biotechnology letters 34 (2012) 663–675.

[81]P. Kanokratana, L. Eurwilaichitr, K. Pootanakit, V. Champreda, Identification of glycosyl hydrolases from a metagenomic library of microflora in sugarcane bagasse collection site and their cooperative action on cellulose degradation, Journal of Bioscience and Bioengineering 119 (2015) 384–391.

[82]A.O.S. Lima, M.C. Quecine, M.H.P. Fungaro, F.D. Andreote, W. Maccheroni, W.L. Araújo, M.C. Silva-Filho, A.A. Pizzirani-Kleiner, J.L. Azevedo, Molecular characterization of a beta-1,4-endoglucanase from an endophytic Bacillus pumilus strain, Applied microbiology and biotechnology 68 (2005) 57–65.

[83]S.J. Kim, J.E. Joo, S.D. Jeon, J.E. Hyeon, S.W. Kim, Y.S. Um, S.O. Han, Enhanced thermostability of mesophilic endoglucanase Z with a high catalytic activity at active temperatures, International journal of biological macromolecules 86 (2016) 269–276.

[84]J. Sakon, D. Irwin, D.B. Wilson, P.A. Karplus, Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca, Nature structural biology 4 (1997) 810–818.

[85]D. Irwin, D.H. Shin, S. Zhang, B.K. Barr, J. Sakon, P.A. Karplus, D.B. Wilson, Roles of the catalytic domain and two cellulose binding domains of Thermomonospora fusca E4 in cellulose hydrolysis, Journal of bacteriology 180 (1998) 1709–1714.

[86]V.V. Zverlov, N. Schantz, W.H. Schwarz, A major new component in the cellulosome of Clostridium thermocellum is a processive endo-beta-1,4-glucanase producing cellotetraose, FEMS microbiology letters 249 (2005) 353–358.

[87]C. Reverbel-Leroy, S. Pages, A. Belaich, J.P. Belaich, C. Tardif, The processive endocellulase CelF, a major component of the Clostridium cellulolyticum cellulosome: Purification and characterization of the recombinant form, Journal of bacteriology 179 (1997) 46–52.

[88]L. Gal, C. Gaudin, A. Belaich, S. Pages, C. Tardif, J.P. Belaich, CelG from Clostridium cellulolyticum: A multidomain endoglucanase acting efficiently on crystalline cellulose, Journal of bacteriology 179 (1997) 6595–6601.

[89]K.-D. Zhang, W. Li, Y.-F. Wang, Y.-L. Zheng, F.-C. Tan, X.-Q. Ma, L.-S. Yao, E.A. Bayer, L.-S. Wang, F.-L. Li, Processive Degradation of Crystalline Cellulose by a Multimodular Endoglucanase via a Wirewalking Mode, Biomacromolecules 19 (2018) 1686–1696.

[90]S. Wu, S. Wu, Processivity and the Mechanisms of Processive Endoglucanases, Applied Biochemistry and Biotechnology 190 (2020) 448–463.

947 [91]F.H. Niesen, H. Berglund, M. Vedadi, The use of differential scanning fluorimetry to detect ligand
948 interactions that promote protein stability, Nature protocols 2 (2007) 2212–2221.

Supplementary Tables

Table S1 CAZymes identified by metagenomic screening on AZCL-HEC or AZCL-xyloglucan

The taxon of the closest homolog, on the order and genus level, was determined by BLASTP similarity searches against the non-redundant protein database.

n.d. the genus of the closest homolog could not be determined.

CAZy family	ID	CBM	presumed activity	% identity	Taxon of closest homolog	
					genus	order
GH3	GH3a	-	beta-glucosidase, beta-xylosidase	66	*Nitrospirillum*	*Rhodospirillales*
GH5	GH5a	-	endoglucanase	60	*Calditrix*	*Calditrichales*
	GH5b	CBM2		68	*Catellatospora*	*Micromonosporales*
GH6	GH6a	CBM2	endoglucanase, exoglucanase	66	*Actinomadura*	*Streptosporangiales*
GH9	GH9a	CBM2 CBM4	endoglucanase, exoglucanase	66	*Sphaerisporangium*	*Streptosporangiales*
	GH9b	-		78	*Thermoflavifilum*	*Chitinophagales*
	GH9c	CBM3		67	*Streptosporangium*	*Streptosporangiales*
	GH9d	CBM2 CBM4		66	*Sphaerisporangium*	*Streptosporangiales*
	GH9e	CBM2 CBM3		83	*Microbispora*	*Streptosporangiales*
	GH9f	CBM4		69	*Sorangium*	*Myxococcales*
GH10	GH10	CBM22 CBM22 CBM22 CBM 9	xylanase	60	*Verrucosispora*	*Micromonosporales*
GH11	GH11a	CBM2	xylanase	62	*Micromonospora*	*Micromonosporales*
	GH11b	-		77	*Thermobifida*	*Streptosporangiales*
GH12	GH12a	-	endoglucanase	70	*Catellatospora*	*Micromonosporales*
GH43	GH43a	-	arabinanase, xylanase	87	***Thermoflavifilum***	*Chitinophagales*
GH48	GH48a	CBM2	endoglucanase, exoglucanase	81	*Microbispora*	*Streptosporangiales*
GH51	GH51a	-	endoglucanase, xylanase	92	***Thermoflavifilum***	*Chitinophagales*
GH53	GH53a	-	galactanase	89	***Thermoflavifilum***	*Chitinophagales*
GH74	GH74a	CBM2	xyloglucanase	92	*Microbispora*	*Streptosporangiales*
	GH74b	CBM2		85	*Microbispora*	*Streptosporangiales*
	GH74c	CBM2		75	*Microbispora*	*Streptosporangiales*
	GH74d	CBM2		81	*Microbispora*	*Streptosporangiales*
	GH74e	-		83	*Microbispora*	*Streptosporangiales*
GH115	GH115a	-	xylane-glucuronidase	64	*Opitutus*	*Opitutales*
CE1	CE1a	-	acetyl xylan esterase	55	*Actinophytocola*	*Pseudonocardiales*
	CE1b	-		57	*Catenulispora*	*Catenulisporales*
	CE1c	-		32	*Lewinella*	*Sphingobacteriales*
CE4	CE4a	-	acetyl xylan esterase	59	n.d.	*Thermomicrobiales*
AA10	AA10a	CBM2	LPMO	64	*Micromonospora*	*Micromonosporales*
CBM2	CBM2a	-	unknown	51	*Dactylosporangium*	*Micromonosporales*
	CBM2b	-	unknown	45	*Micromonospora*	*Micromonosporales*

Table S2 Distance matrix of the multiple sequence alignment of 174 characterized GH9s

(see Excel file)

Table S3 Protein domains attached to at least one of the 174 characterized GH9s and their occurrence

Domain	Pfam accession number	Occurrence	
		Protein number	%
Glycoside hydrolase family 9	PF00759	174	100,0
Cellulase N-terminal Ig-like domain	PF02927	55*	33,3
CBM3	PF00942	47	27,0
Dockerin	PF00404	41	23,6
CBM2	PF00553	18	10,3
CBM4/9	PF02018	16	9,2
Fibronectin type III domain	PF00041	6	3,4
CBM49	PF09478	4	2,3
Glycoside hydrolase family 48	PF02011	3	1,7
PKD domain	PF00801	2	1,1
CBM10	PF02013	2	1,1
CBMX2	PF03442	2	1,1
CBM64	PF18666	2	1,1
Glycoside hydrolase family 5	PF00150	1	0,6
Glycoside hydrolase family 16	PF00722	1	0,6
F5/8 type C domain	PF00754	1	0,6
Carbohydrate family 9 binding domain-like	PF06452	1	0,6
Immunoglobulin I-set domain	PF07679	1	0,6
Glycoside hydrolase family 44	PF12891	1	0,6

* 55 Ig-like domain-containing proteins according to Pfam database; 3 supplementary proteins are identified respectively with CDD database (UniProt References C9RJA3 and Q6LUT2) and InterPro (Q9APG3)

Supplementary figures

Figure S1: Alignment of 168 characterized GH9 enzymes and the six identified metagenomic GH9 enzymes.

The multiple sequence alignment was built using Clustal Omega, and the aligned fasta file was submitted to ESPript 3.0 (http://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi). Identical or similar amino acids are colored in red, and boxed in blue if the global similarity score is > 0.7. 100% conserved amino acids are highlighted in red. Loop A of Ig-like-domaine-containing enzymes is located between positions 467 and 483. Residues composing loop B are situated between positions 453 and 454.

(see corresponding pdf file)

Figure S2: Superposition of the three-dimensional structures of three bacterial GH9s belonging to distinct clades

X-ray crystal structures of *T. fusca* Cel9A (Sakon et al. 1997) (UniProt Reference P26221, PDB ID: 1TF4, green), *C. thermocellum* CbhA (Schubot et al. 2004) (Q6RSN8, PDB ID: 1UT9, blue) and *R. cellulolyticum* Cel9M (Parsiegla et al. 2002) (Q9EYQ2, PDB ID: 1IA6, red) were superposed using PyMol (DeLano 2002). Protein three-dimensional structures are displayed using the cartoon representation.

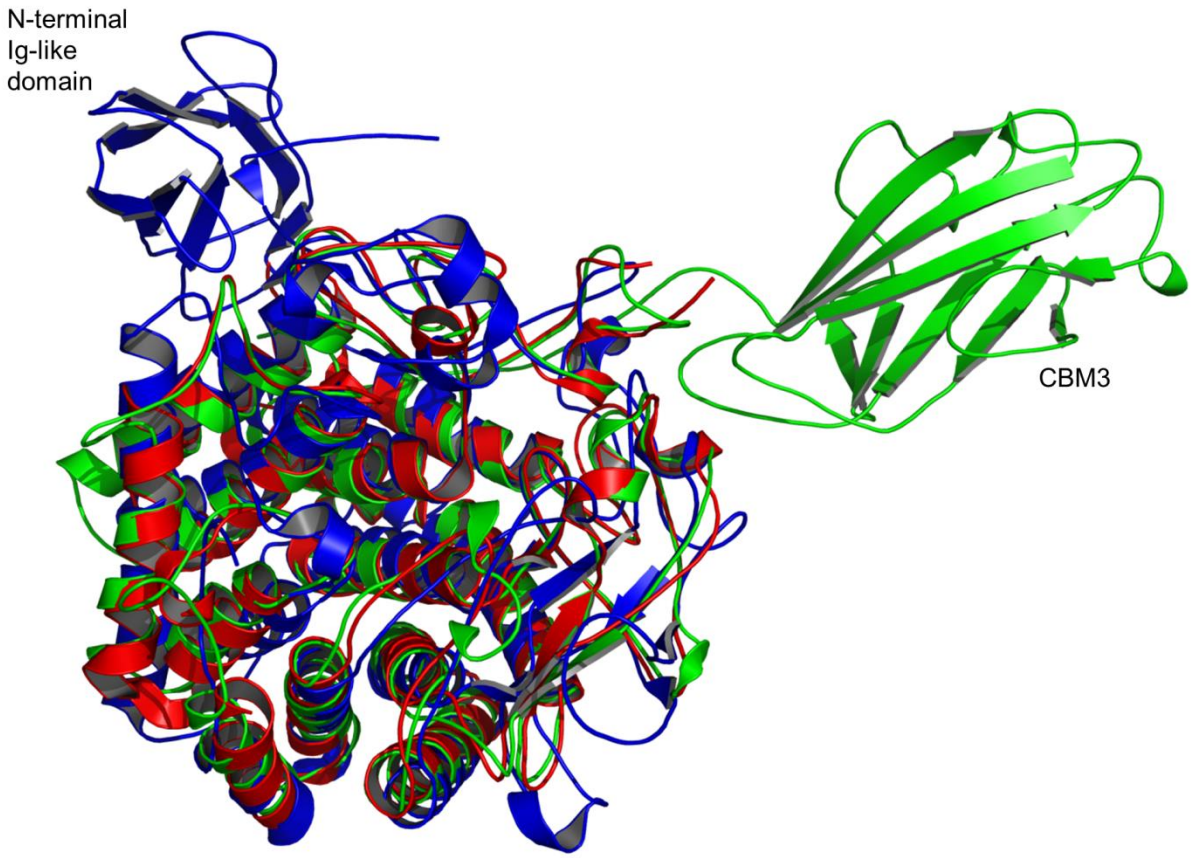


Figure S3: Superposition of the three-dimensional structures of two bacterial GH9s belonging to distinct clades with co-crystallized substrate

X-ray crystal structures of *T. fusca* Cel9A (Sakon et al. 1997) (UniProt Reference P26221, PDB ID: 1JS4, green), *C. thermocellum* CbhA (Schubot et al. 2004) (Q6RSN8, PDB ID: 1RQ5, red) were superposed using PyMol (DeLano 2002). Protein three-dimensional structures are displayed using the cartoon representation. Co-crystallized substrates (cellotriose for TfCel9A and cellotetraose for CtCbhA) are displayed as ball-and-stick. Loops A (Arg558-Gln575, CtCbhA) and B (Trp430-Ser436, TfCel9A) are respectively colored in magenta and red.

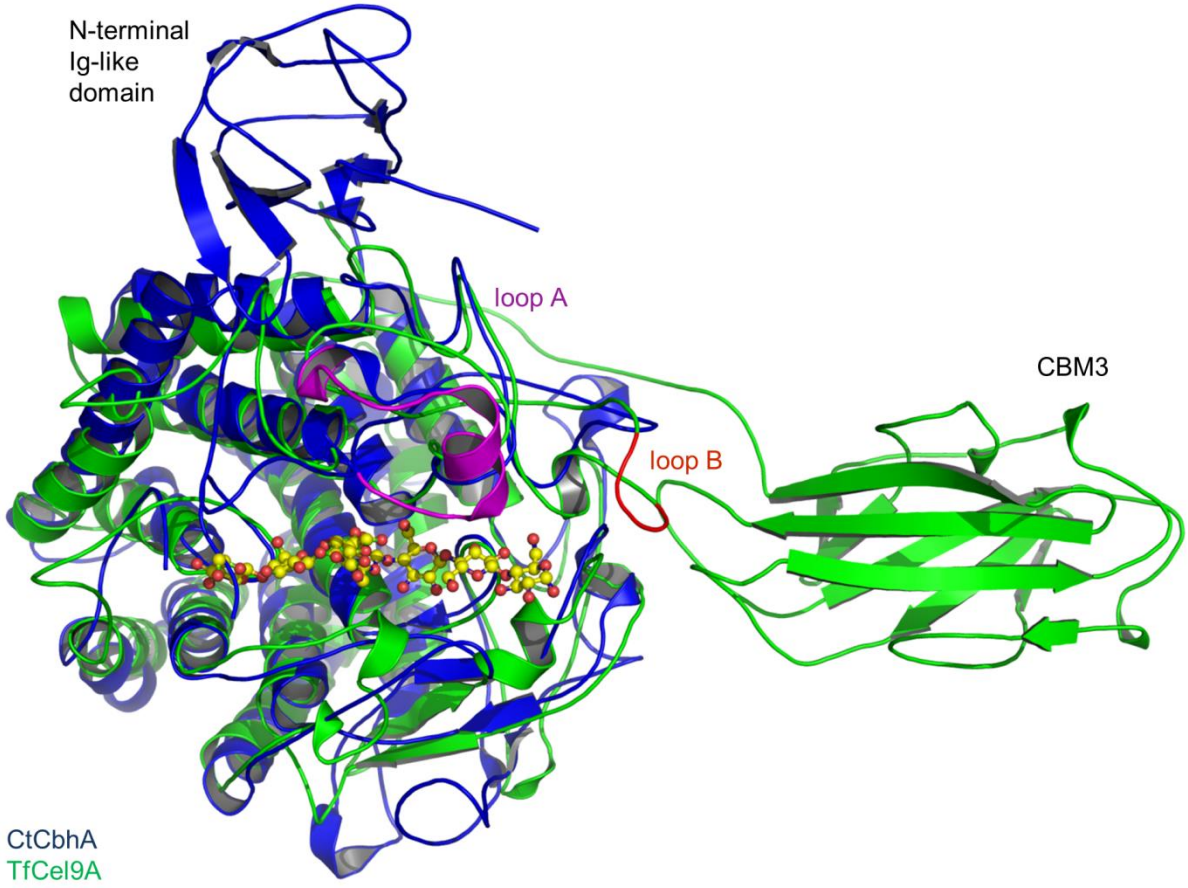


Fig. S4: Purification of recombinant metagenome-derived bacterial GH9s

(A) SDS-PAGE of recombinant GH9 enzymes purified by affinity chromatography on a nickel chelating resin (GH9e) followed by gel filtration (GH9b and GH9d). GH9b (65.8 kDa), GH9d (93.5 kDa) and GH9e (83.5 kDa) were purified from the 12 000 x *g* supernatant of the lysates of *E. coli* cultures and 0.5 µg were separated by SDS-PAGE. (B) Gel filtration chromatography of recombinant GH9 enzymes. Purified GH9b, GH9d and GH9e (respectively 1.6 mg, 3.2 mg and 0.3 mg) were loaded on a Superdex 200 10/300 GL (GE Healthcare) using 25 mM triethanolamine pH 7.0, 150 mM NaCl, 5% (v/v) glycerol, as buffer. For calibration of the column, the following molecular weight markers (BioRad, pointed by black diamonds) were used: thyroglobulin (670 kDa), γ-globulin (158 kDa), ovalbumin (44 kDa), myoglobin (17 kDa) and vitamin B12 (1.35 kDa). The elution was monitored by measuring the absorbance at 280 nm. GH9b (solid line), GH9d (dotted line) and GH9e (dashed line) main peak were eluted at respective apparent molecular weights of 28 kDa, 287 kDa and 40 kDa. MW, molecular weight

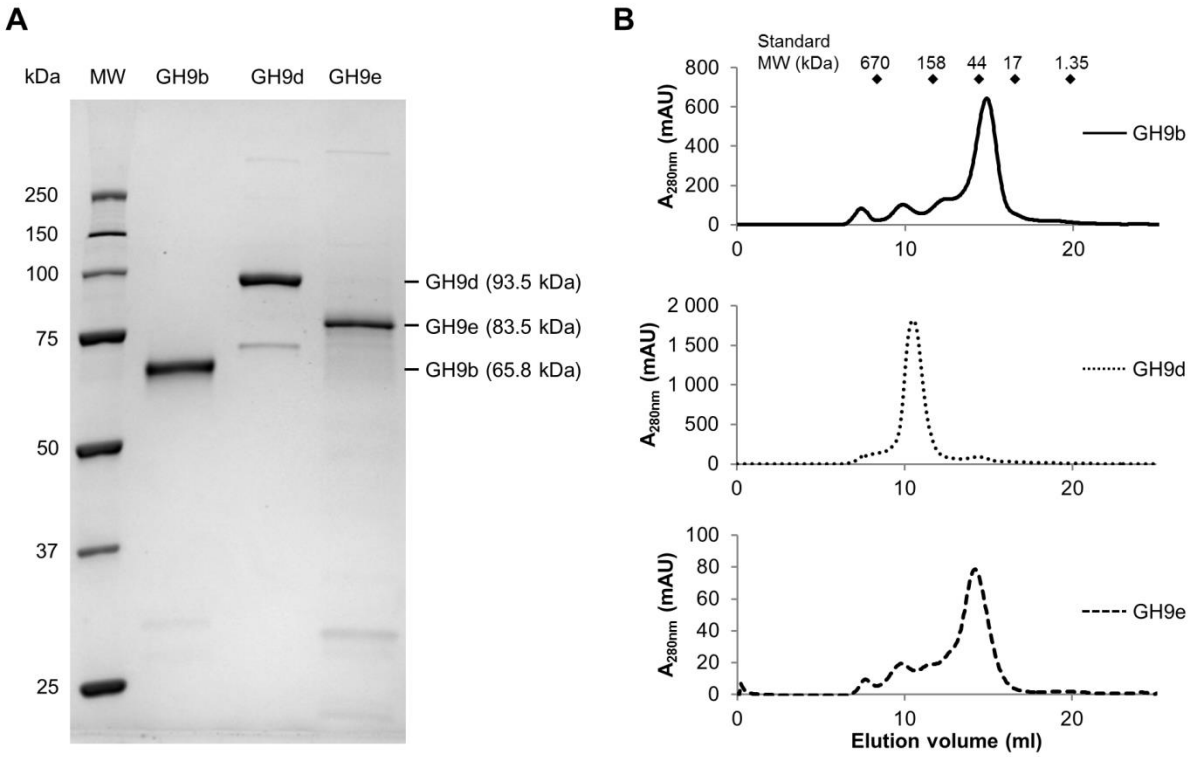


Figure S5: Purification and domain architecture of the GH48 enzyme retrieved from the metagenome

Right, schematic representation of the mature protein are presented according to its relative sequence length. Mature protein molecular weight was calculated using ExPASy ProtParam tool (Gasteiger et al., 2005). Taxonomic order was predicted by BLAST using the closest homolog corresponding organims (Madeira et al. 2019). N, N-terminal, C. C-terminal. Left, SDS-PAGE of purified recombinant GH48 produced in *E. coli*.