Sequence, Structure, and Evolution of Cellulases in the Glycoside Hydrolase Family 48

Leonid O. Sukharnikov[a,1,2], Markus Alahuhta [a,1,3], Roman Brunecky [1,3], Amit Upadhyay[1,2], Michael E. Himmel[1,3], Vladimir V. Lunin[1,3,b] and Igor B. Zhulin[1,2,b]

[1]BioEnergy Science Center and [2]Joint Institute for Computational Sciences, University of Tennessee – Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[3]Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401, USA

Running title: *Defining cellulase in the glycoside hydrolase family 48*

[a]L.O.S. and M.A. contributed equally to this work

[b]To whom correspondence should be addressed: Igor B. Zhulin, Email: joulineib@ornl.gov or Vladimir V. Lunin, Email: vladimir.lunin@nrel.gov

Keywords: cellulase; horizontal gene transfer; structure; metagenome

Background: Cellulases are non-homologous isofunctional enzymes, which prevents their unambiguous identification in genomic datasets.
Results: Cellulases from glycoside hydrolase family 48 have distinct evolutionarily conserved sequence and structural features.
Conclusion: Conserved sequence/structure features can be used to differentiate cellulases from non-cellulases in genomic datasets.
Significance: Unambiguous identification of cellulases in genomic data is critical in searching for novel cellulolytic activities needed for bioenergy research.

SUMMARY

Currently, the cost of cellulase enzymes remains a key economic impediment for commercialization of biofuels (1). Enzymes from glycoside hydrolase family 48 (GH48) are a critical component of numerous natural lignocellulose degrading systems. Although computational mining of large genomic datasets is a promising new approach for identifying novel cellulolytic activities, current computational methods are unable to distinguish between cellulases and enzymes with different substrate specificities that belong to the same protein family. We show that by using a robust computational approach supported by experimental studies, cellulases and non-cellulases can be effectively identified within a given protein family. Phylogenetic analysis of GH48 showed non-monophyletic distribution, an indication of horizontal gene transfer (HGT). Enzymatic function of GH48 proteins coded by horizontally transferred genes was verified experimentally, which confirmed that these proteins are cellulases. Computational and structural studies of GH48 enzymes identified structural elements that define cellulases and can be used to computationally distinguish them from non-cellulases. We propose that the structural element that can be used for *in silico* discrimination between cellulases and non-cellualses belonging to GH48 is an omega loop located on the surface of the molecule and characterized by highly conserved rare amino acids. These markers were used to screen metagenomics data for 'true' cellulases.

INTRODUCTION

The recent exponential growth of genomic data presents a unique opportunity to search for novel cellulolytic activities. However, the absence of a clear understanding of structural and functional features that are critical for decisive computational identification of cellulases prevents their identification in these datasets. True cellulases are defined as enzymes that show biochemical activity on cellulose substrates (i.e. crystalline or amorphous cellulose). Strikingly, all known cellulases have homologs that have similar protein folds and even amino acid sequences, but do not show biochemical activity on cellulosic substrates (2), which makes computational-only identification of true cellulases error-prone. Glycoside hydrolase family 48 (GH48) is one of the many families defined in the CAZy (Carbohydrate-Active EnZymes) database(3) that contains biochemically confirmed cellulases. Furthermore, GH48 cellulases are considered the key component of various cellulolytic systems (4-6). They are highly expressed in cellulolytic bacteria, such as *Clostridium cellulolyticum, C. cellulovorans, C. josui and C. thermocellum* and many others *(4)*. In *C. thermocellum*, a bacterium that exhibits one of the highest rates of cellulose degradation among all known cellulolytic bacteria, GH48 cellulases are up-regulated during growth on crystalline cellulose (4). Hence, these enzymes become the most abundant subunits in the *C. thermocellum* cellulosome, a complex of enzymes highly efficient in cellulose degradation (4,7). Notably, complete knockout of both GH48 enzymes in *C. thermocellum* leads to a significant decrease in performance, but does not completely abolish cellulolytic activity (4), whereas knockout of the GH48 gene in *Ruminococcus albus* (5) leads to nearly complete loss of cellulase activity.

Usually, only one (rarely two or three) gene(s) encoding GH48 enzymes can be found in the genomes of cellulose-degrading bacteria (6), whereas genes for GH5 and GH9 cellulases are present in much higher numbers (8,9). Interestingly, GH48 cellulases often act in synergy with GH9 cellulases, which increases their catalytic activity dramatically (10), a feature that may be utilized for industrial application of these enzymes; for example, "designer cellulosomes" (11)

Experimental studies revealed that some GH48 cellulases have only cellulolytic activity, and thus cannot hydrolyze other substrates (i.e., xylan and mannan) (12). A few GH48 cellulases have mixed substrate specificity, for example, are capable of degradation of xylan (13) or β-glucan (14) in addition to cellulose. There are two GH48 enzymes from beetles *Gastrophysa atrocyanea* that are unable to hydrolyze cellulose-containing substrates (i.e., Avicel, carboxymethylcellulose, acid swollen cellulose, etc.), whereas they showed distinct enzymatic activity toward chitin (15) (supplemental Table S1).

Previous genomic studies have shown that GH48 enzymes are found in fungi; as well as in bacteria, including Clostridia, Bacilli (both Firmicutes) and Actinobacteria. However, the presence of the GH48 cellulase (16) in the evolutionarily distant deltaproteobacterium, *Myxobacter sp.* AL-1, was never explained.

Here we report evolutionary studies of GH48 enzymes, present a crystal structure of the GH48 enzyme encoded by a horizontally transferred gene, and characterize structural and functional differences between cellulases and chitinases in this group of enzymes. We also show that our computational approach can be used to search for true GH48 cellulases in metagenomic databases.

EXPERIMENTAL PROCEDURES

Bioinformatics software and computer programming environment

The following software packages were used in this study: HMMER v3.0 (17), MAFFT v6.0 (18), MEGA v5.0 (19), Jalview v2.7.0 (20), BLAST v2.2.17 (21). All multiple sequence alignments were built in MAFFT with its L-INS-i algorithm. All maximum likelihood phylogenetic trees were built in PhyML (22) with LG + Γ_4 + F parameters. Symmetrical best hits (SymBets) were assigned using BLAST algorithm.

All computational analyses were performed in a local computing environment, and custom scripts for data analysis were written in BioPerl. A remote version of the NCBI non-redundant database was used for direct queries using BioPerl scripts. A local version of the same database was used for querying with the *hmmsearch* algorithm of the HMMER package.

Data sources and literature analysis - National Center for Biotechnology Information

(NCBI) non-redundant database (nr) in FASTA format as of April 2011 was retrieved from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/. Hidden markov Model (HMM) of glycol_hydro_48 (PF02011) was retrieved from the Pfam database vPfam26 (23). Structures of Cel48S from *C. thermocellum* (24) and Cel48F from *C. cellulolyticum* (25) were retrieved from RCSB PDB: www.pdb.org (26) .

Glycoside hydrolases of family 48 (classification according to CAZy database (3)) with known activity were identified from the literature (supplemental Table S1) and then mapped on the phylogenetic tree of GH48 enzymes in order to place the functional knowledge into the taxonomic context. Enzymes were considered of demonstrated cellulolytic function if their activity had been analyzed by *in vitro* biochemical studies.

Multiple sequence alignment and construction of phylogenetic tree

183 GH48 protein sequences were retrieved from the NCBI nr database using *hmmsearch* of the HMMER package (17) with Pfam gathering threshold and Pfam domain model glycol_hydro_48 (>600 amino acid residues). Then GH48 enzymatic domains corresponding to the Pfam model were excised from the protein sequences using BioPerl scripts and further analyzed. 69 domain sequences were found to be too short (<300 amino acid residues) and thus discarded to improve quality of the subsequent studies. 114 GH48 sequences were taken into further analysis.

Multiple sequence alignment (MSA) of 114 GH48 domains was constructed in MAFFT. The resulting alignment was used to build a maximum likelihood tree in PhyML. The conservation pattern in the MSA was analyzed in Jalview (20) with underlying tools, and phylogenetic tree was analyzed using the MEGA5 package. Taxonomy assignments for the proteins on the tree were taken from GenPept records from NCBI protein database.

Identification of orthologs, paralogs and horizontally transferred genes

Because GH48 is typically present as one copy per genome we assigned as orthologs all GH48 protein sequences that: (i) form a monophyletic clade on the tree; (ii) were present as a single copy per genome; (iii) come from phyla with the same common ancestor after species-proteins tree topology reconciliation: Firmicutes, Actinobacteria and Chloroflexi; (iv) were characterized by symmetrical best matches (SymBets). Similar GH48 sequences that were present in two or more copies per genome were assigned as paralogs.

Horizontally transferred genes were defined in two ways: (i) by means of phylogenetic studies, where HTGs were assigned based on phyletic distribution on the tree (27) (ii) by means of probabilistic approach (27), where probability of occurrence of GH48 genes in prokaryotic genomes was calculated as the percentage of genomes containing GH48 genes divided by the total number of the available genomes, assuming that each genome contains only one GH48 gene (Table 1).

Metagenomic data analysis

We analyzed a publicly available dataset of protein sequences from microbial communities in 295 metagenome samples retrieved from JGI/M (28) as of October, 2011 and Cow Rumen dataset from (29). Sequences encoding glycoside hydrolase family 48 proteins [glycol_hydro_48 (PF02011) Pfam domain model] were collected from metagenome datasets with *hmmsearch*. Duplicate sequences were removed.

Cloning, expression and purification of H. chejuensis GH48

A codon optimized pMAL expression plasmid obtained from DNA 2.0 (DNA2.0 Menlo Park, CA) containing the *H. chejuensis* catalytic domain was transformed into *E. coli* (BL21) (Agilent, Santa Clara, CA), and overexpresed 37 °C with the presence of 0.3 mM IPTG. The recombinant fusion protein contained a *C*-terminal maltose binding domain and was purified using a amylose high flow resin (NEB, Ipswich, MA) the eluted fusion protein was then cleaved using a genenase protease site incorporated into the sequence (NEB, Ipswich, MA). The *H. chejuensis* GH48 module was further purified by anion chromatography on a source 15Q column (GE Healthcare, Piscataway, NJ). Using buffers A (20 mM Tris, pH 6.8) and B (20 mM Tris, pH

6.8 2M NaCl). Minor impurities were removed by size exclusion chromatography using HiLoad Superdex 75 (26/60) (GE Healthcare) in 20 mM acetate buffer, pH 5.0 containing 100 mM NaCl and 1 mM sodium azide. The purified protein solution was concentrated with a Vivaspin 5K concentrator (Vivaproducts, Littleton, MA), and its concentration was measured using BCA assay (Pierce, Rockford, IL).

Model substrate and pretreated biomass

Avicel (PH101), and phosphoric acid swollen cellulose (PASC) generated from Avicel, were used to evaluate the cellulolytic efficiency of *H. chejuensis* GH48. To provide a basis for the maximum theoretical sugar yield achievable from each substrate during enzymatic hydrolysis, portions of each of the pretreated solids samples were dried and subjected to the standard two-stage sulfuric acid hydrolysis method for determining structural carbohydrates in lignocelluloses as described by Sluiter and coworkers (30). In this method, the carbohydrate content of each pretreated sample is calculated from the carbohydrates released. In both cases, it is approximately 95% glucan.

Enzymatic Digestion Assays

GH48 activity was determined at 45ºC, at an enzyme loading of 15 mg/g glucan Avicel or 80 mg/g glucan PASC in 20 mM acetate buffer pH 5.5, containing 10 mM $CaCl_2$ and 100 mM NaCl. The assay slurry was mixed by inversion. Digestions were run continuously for 7 d and sugar release was monitored by removing aliquots. Samples taken at various time points and the enzymes were inactivated by boiling for 15 min. Samples were then filtered through 0.45 µm Acrodisc syringe filters and analyzed for glucose and cellobiose by HPLC. Samples were injected at 20 µL and run on an Agilent 1100 HPLC system equipped with an BioRad Aminex HPX-87H 300 mm x 7.8 mm column heated to 55^oC. A constant flow of 0.6 mL/min was used with 0.1M H_2SO_4 in water as the mobile phase to give separation of the analytes. Glucose, and cellobiose were quantified against independent standard curves. All experiments were performed in triplicate and the resulting extents of conversion are shown as percent glucan converted.

CD methods

CD measurements were carried out using a Jasco J-715 spectropolarimeter with a jacketed quartz cell with a 1.0 mm path length. The cell temperature was controlled to within +/− 0.1^oC by circulating 90% ethylene glycol using a Neslab R-111m water bath (NESLAB Instruments, Portsmouth, NH) through the CD cell jacket. The results were expressed as mean residue ellipticity $[e]_{mrw}$. The spectra obtained were averages of five scans. The spectra were smoothed using an internal algorithm in the Jasco software package, J-715 for Windows. Protein samples were studied in 20 mM sodium acetate buffer, pH 5.0 with 100 mM NaCl at a protein concentration of 0.25 mg/mL for the near-UV CD. Thermal denaturation of different constructs was monitored by CD in the near UV (190–260 nm) region. For the analysis of thermal stability, the temperature was increased from 25 to 60^oC with a step size of 0.2^o C, and monitored at a wavelength of 222 nm.

Crystallization

H. chejuensis GH48 (YP_433697) crystals were obtained with sitting drop vapor diffusion using a 96-well plate with Crystal Screen HT from Hampton Research (Aliso Viejo, CA). Fifty µL of well solution was added to the reservoir and drops were made with 0.2 µL of well solution and 0.2 µL of protein solution using a Phoenix crystallization robot (Art Robbins Instruments, Sunnyvale, CA). The crystals were grown at 20°C with 0.05 M potassium phosphate monobasic and 20% w/v polyethylene glycol 8000 as the well solution. The protein solution contained 15 mg/mL of protein, 20 mM acetic acid pH 5, 100 mM NaCL and 10 mM $CaCl_2$.

Data collection and processing

The *H. chejuensis* crystal was flash frozen in a nitrogen gas stream at 100 K before home source data collection using a Bruker X8 MicroStar X-Ray generator with Helios mirrors and a Bruker Platinum 135 CCD detector. Data were indexed and processed with the Bruker Suite of programs version 2011.2-0 (Bruker AXS, Madison, WI).

Structure solution and refinement

Intensities were converted into structure factors and 5% of the reflections were flagged for Rfree calculations using programs F2MTZ, Truncate, CAD and Unique from the CCP4 package of programs (31). The GH48 structure was solved using molecular replacement with program Molrep (32) with PDB entry 1G9G as a model. ARP/wARP (33) version 7.0 and Coot (34) version 0.6.2 was used for multiple cycles of automatic and manual model building. Further refinement and manual correction was performed using REFMAC5 (35) version 5.6.0117 and Coot. The MOLPROBITY method (36) was used to analyze the Ramachandran plot and root mean square deviations (rmsd) of bond lengths and angles were calculated from ideal values of Engh and Huber stereochemical parameters (37). Wilson B-factor was calculated using CTRUNCATE (31) version 1.0.11. Average B-factors, were calculated using program ICM version 3.7-2a (Molsoft LLC, La Jolla, CA). The resulting structures have been deposited to the Protein Data Bank with PDB code 4FUS. The data collection and refinement statistics are shown in supplemental Table S2.

RESULTS

Phyletic distribution of GH48 sequences and horizontal gene transfer

GH48 enzymes that were retrieved from databases belong to only four prokaryotic phyla (Actinobacteria, Firmicutes, Chloroflexi, and Proteobacteria) and only two eukaryotic phyla (Fungi and Arthropoda) indicating a rather unusual evolutionary history. Taking into account that Firmicutes, Actinobacteria and Chloroflexi (i) likely shared a common ancestor (38), (ii) showed GH48 enrichment compared to other phyla (Table 1), and (iii) contained a significant number of biochemically confirmed GH48 cellulases while lacking any confirmed non-cellulases, we hypothesize that the GH48 cellulase originated in the last common ancestor of Firmicutes, Actinobacteria and Chloroflexi. Therefore, we have first analyzed sequences only from these three phyla that satisfied two additional criteria: (i) were present as the only GH48 gene in a genome and (ii) showed many-to-many symmetrical best hits (SymBets) relationships (39) As a result, 65 sequences, which included 12 biochemically confirmed cellulases, were taken into further analysis and aligned. The maximum-likelihood tree constructed from this alignment was monophyletic, i.e. sequences from the same phylum were found in a single clade. In the next step, we determined the conserved residues in the alignment and found that all functionally important sites (including folding and substrate binding) were invariably conserved (supplemental Table S3).

Because paralogs typically have a similar, but not identical function, we asked a question whether paralogous GH48 sequences may represent enzymes with different substrate specificity. If so, they should show differences in some of the highly conserved sites, especially those implicated in substrate binding. Surprisingly, we found that paralogous GH48 sequences in genomes of Firmicutes and Actinobacteria were nearly identical (90-98% identity) and retained all conserved residues that were identified in the set of orthologous sequences. It appears that the functional innovation in paralogs resides not in the catalytic domain, but in the repertoire of their auxiliary domains (Fig. 1).

The evidence of horizontal gene transfer emerges when a protein sequence from a particular organism shows high similarity to a homolog from a distant taxon (27). In case of GH48, all sequences from Fungi were found in the middle of the Firmicutes clade, whereas all sequences from Insecta were found in the middle of the Actinobacterial clade (Fig. 2).This non-monophyletic distribution clearly suggests horizontal gene transfer into eukaryotes from the two prokaryotic phyla.

Thus, a total of 23 horizontally transferred genes were identified through phylogenomic analysis, where an implicitly defined (see above) set of orthologs showed the presence of non-monophyletic clades with representatives of Proteobacteria, Fungi and Insecta (Fig. 2). Additionally, in prokaryotes, they were also identified by a probabilistic approach (27), where relative increases in abundance of GH48 genes in the genomes of Actinobacteria, Firmicutes and Chloroflexi was compared to that of Proteobacteria, as described in materials and methods (Table 1). Notably, Actinobacteria, Firmicutes and Chloroflexi genomes had much higher probability of occurrence of GH48 genes compared to Proteobacteria, Fungi and insects, which along with their distribution on the phylogenetic tree presents additional evidence for horizontal gene transfer into latter.

In summary, here we define all GH48 orthologs and paralogs from Actinobacteria, Firmicutes and Chloroflexi as true cellulases based on phylogenomic analysis, which correlates with their experimentally confirmed enzymatic activities (Fig. 2; supplemental Table S1).

Cellulose digestion by the horizontally transferred GH48

A comprehensive list of all biochemically studied GH48 cellulases is presented in the supplement (supplemental Table S1). This list shows that previously studied cellulases are mostly present in Firmicutes and Actinobacteria with a single representative of proteobacteria (*Myxobacter* sp. Al-1). We determined the activity of the GH48 enzyme from a proteobacterium *Hahella chejuensis,* which was a subject of horizontal gene transfer, on both crystalline and amorphous substrates (supplemental Fig. S1). These results showed that *H. chejuensis* is a cellulase, because it shows activity on the PASC substrate. The poor performance on the more crystalline substrate is likely due to the lack of the CBM domains in our construct, which is critical for optimal performance on a crystalline substrate, such as Avicel.

Crystal Structure of H. chejuensis GH48

The structure of HcheGH48 was refined to a resolution of 1.75 Å with R and R_{free} of 0.154 and 0.205, respectively. There is one molecule in the asymmetric unit in complex with a cellobiose molecule bound at the product position. It has an $(alpha/alpha)_6$ barrel fold with one calcium and two sodium atoms; and multiple ethylene glycol, glycerol, acetate and phosphate molecules. Due to the long crystallization time (more than one year), two residue modifications were observed: a 2-oxohistidine at position 352 and polyethylene glycol modification of Tyr439. There were two outliers in the Ramachandran plot, Glu72 and Ala73, both of them well defined by the density and close to the allowed region.

Structural comparison to other known GH48s

Pair wise secondary-structure matching of structures with at least 70% secondary structure similarity by PDBefold (40) found 22 unique structural matches for HcheGH48 from the protein data bank. All similar structures were CelF, CelS or CelA GH48 variants with secondary structure similarity between 79% and 88%. Closer inspection of the structure shows that the overall fold (Fig. 3) and the catalytic tunnel are almost identical to *C. cellulolyticum* CelF, *C. thermocellum* CelS and *Caldicellulosiraptor bescii* CelA. In HcheGH48 Glu83 is the catalytic residue. The residues lining the tunnel, catalytic Glu83 and the positions of the sugar rings of the cellobiose molecule are mostly conserved when compared to the *Clostridium cellulolyticum* CelF, *Clostridium thermocellum* CelS and *C. bescii* CelA GH48 structures (pdb codes 1FCE (41), 1L2A (24) and 4EL8). The identical residues lining the pocket are Trp344, Gln247, Asp241, Ser245, Thr239, Ser136, Phe346, Lys303, Tyr331, Thr251, Gln207, Phe206, Asn204, Trp180, Trp330, Tyr357, Trp450, Trp453, Trp447, His64, Arg648, Trp650, Asp529, Glu83 and Glu83. The biggest differences are Trp450 and Ala616. Trp450 is a methionine in CelF GH48 and phenylalanine in *Clostridium thermocellum* CelS and *C. bescii* CelA GH48s. Ala616 is a histidine in CelS GH48 and alanine in the other structures.

Closer inspection of the omega loop shows that it is defined by two anchor residues, Trp508 and Asn516 (Fig. 3). Comparison to *C. cellulolyticum* CelF, *C. thermocellum* CelS and *C. bescii* CelA GH48 structures shows that these residues are conserved and have identical conformation in all four structures. The omega loop of HcheGH48 differs from the others by having a proline at position 523, causing a local conformational change, where the other structures have a tyrosine which further anchors the loop. This, however, does not change the overall conformation or position of the loop but does suggest that variability in the loop is possible without affecting activity.

Conserved amino acid positions in the GH48 family in the context of structure

We used sequence numbering of Cel48F from *C. cellulolyticum* H10 to designate amino acids in all multiple sequence alignment studies, because it is the most extensively studied GH48 structure currently available (25,41,42). Literature and MSA analysis showed that all GH48 enzymes have 100% conserved catalytic acid and base positions (E55 and D230 in Cel48F) thus these residues are not discussed.

There are three major types of amino acids that participate in substrate recognition and correct folding of the GH48 enzymes: hydrophobic stacking interactions, hydrogen bonding and calcium coordination residues (supplemental Table S3) (24,25,41,42). All these residues are highly conserved in orthologs from Actinobacteria, Firmicutes and Chloroflexi; as well as in Proteobacteria, which indicates that genes horizontally transferred to Proteobacteria code for cellulases, a statement confirmed biochemically (this work and (16)). Our results also revealed that the GH48 enzyme from *H. chejuensis* does not possess any additional elements that would differentiate it from other cellulases.

Consequently, we hypothesize that fungal GH48s are also cellulases due to their high sequence similarities with cellulolytic orthologs, and the fact that almost all important for catalysis residues (supplemental Table S3) are highly conserved in fungi (supplemental Table S4) with only one exception – the Ca^{2+} coordination residues, which were considered to play role in thermal stability of GH48 enzymes (24), but not in substrate specificity. In contrast, GH48 enzymes from all insects are represented by non-cellulases, because of large number of amino acid substitutions in positions that are conserved among cellulases, one omega loop deletion, and lack of cellulolytic activity confirmed biochemically (15).

Mutations in critical positions were not found in all sequences from insects (supplemental Table S4).Thus, MSA and structural analyses suggested that the major difference between cellulases and non-cellulases (i.e., chitinases) from insects is the additional omega loop located between P469 and A482 (as in Cel48F) in all cellulases. This omega loop includes two residues highly conserved in all cellulolytic orthologs (99-100% conservation): W472 and N481. Residue L484 (as in Cel48F), located adjacent to the loop and strictly conserved in cellulolytic orthologs is also mutated in all insects. This omega loop located on the surface of GH48 molecule and connects two beta-strands that form one side of the catalytic tunnel near the exit of the product (Fig. 3). Thus, here we report structural differences that occurred after an event of horizontal gene transfer from Actinobacteria to Insecta that caused mutation of cellulases to chitinases.

Screening metagenomic datasets for GH48 cellulases

Sequences of 211 GH48 proteins were retrieved from the combined metagenome dataset (>79 million sequences) with hmmsearch of HMMER (17) and glycol_hydro_48 Pfam domain model with the Pfam gathering threshold. Then 36 duplicates were removed and remaining 175 sequences were used in protein BLAST queries. BLAST results showed that these sequences belong to the same major phyla as sequences belonging to well-defined genomes that were retrieved from the NCBI nr database: Actinobacteria, Firmicutes, Chloroflexi, Proteobacteria, and insects (Arthropoda), except for Fungi (Fig. 4). These results indicate that fungal species are either absent from the metagenomes used in this study or significantly underrepresented. In summary, 9 sequences from metagenomics samples belonged to insects and were classified as non-cellulases and the other 166 sequences were classified as cellulases, based on the phylogenomic and structural evidence presented above.

To confirm the validity of this classification, 166 metagenomic GH48 sequences classified as cellualses were aligned by hmmalign of HMMER (17) with default Pfam parameters. MSA analysis (supplemental Table S5) showed that 93% of sequences have all of the residues important for protein folding and catalysis with very few conservative substitutions that were also found in some of the cellulolytic orthologs from complete genomes. A few non-conservative substitutions that were found in a small set of the sequences (~7% of all) could indicate potential differences in function or could simply be sequencing/assembly errors, a rather common problem in metagenomics (43,44) Therefore, experimental evidence must be obtained to clarify this point.

Because metagenomic samples show a large variation in the total number of genes sequenced (for example, a wastewater treatment plant metagenome has 30169 genes, whereas a biofuel metagenome has 2706009 genes), the percentage of GH48 domains in each metagenome was calculated (Fig. 5). These metagenomes were also grouped together according to their habitats and the percentage abundance of GH48 in each habitat was also calculated (Fig. 5).

DISCUSSION

Using a phylogenomic approach, we have determined that the GH48-type β-1,4- glycoside hydrolyses might have originated from a common ancestor of three closely related phyla – Firmicutes, Actinobacteria and Chloroflexi (38). We have determined a number of gene duplication events in representatives of these phyla and several cases of horizontal gene transfer. For example, fungi received these genes horizontally from a representative of Firmicutes, whereas insects received these genes from a representative of Actinobacteria. Similarly, representatives of Proteobacteria also received their GH48 genes horizontally. By comparing orthologous sequences from Firmicutes, Actinobacteria and Chloflexi we identified a number of amino acid positions that are uniquely conserved in this group of organisms. Satisfactorily, the only activity that was previously found in this group is that of a cellulase. Thus, we suggest that conserved positions in the catalytic domains from Firmicutes, Actinobacteria and Chloflexi can be used as genomic signature for a GH48 cellulase.

We then wondered if this genomic signature for a cellulase remains intact in paralogs and horizontally transferred genes, because these types of genes often assume a slightly different function. For example, just one or a few mutations in a catalytic domain may lead to different substrate specificity. Notably, screening and studying paralogous sequences of GH48 proteins showed no significant differences in their catalytic domains, but rather noticeable differences in their auxiliary domains (i.e., cellulose binding domain (CBM), fibronectin type III-like domain and etc.). On the contrary, genes that were horizontally transferred from Actinobacteria to insects (Metazoa) acquired a new activity to hydrolyze chitin, but lost the ability to degrade cellulose.

Following this initial evolutionary analysis, we extended our findings to structural analysis of GH48 enzymes. We found that all orthologs and paralogs have a 10 to 14 residue omega loop (P469-A482 as in Cel48F), that has no counterpart in enzymes from insects. Moreover, this omega loop is constituted by highly conserved amino acids (W472 and N481 as in Cel48F) and located on the surface of the molecule. Thus, in accord with classical definition of omega loops (45), it may play the following roles in this enzyme structure: folding, stability or it may contribute to the dynamics of the enzyme during catalysis.

High conservation of the omega loop residues in cellulases suggests its importance for the computational identification of cellulases and complete absence of the loop in all non-cellulases indicates that GH48 chitinases lost this structural element. We hypothesize that the absence of the loop in chitinases allows more conformational degrees of freedom in the active site tunnel upon binding of the substrate, which permits a bulkier chitin 'slide' freely. In contrast, cellulases may have more rigid structures 'reinforced' by the omega loop. Regardless of the exact role of the omega loop, which can be determined only experimentally, we suggested that it is important for cellulolytic activity, which allowed us to design a strategy to identify new cellulases in metagenomic data.

Thus, phylogenomic and structural analyses of GH48 suggest that proteins from Actinobacteria, Firmicutes, Chloflexi and Proteobacteria are indeed cellulases. Biochemical activities of GH48 proteins from two Pyromyces species have never been studied; thus it is unknown whether they are cellulases. However, because these proteins are not only homologous to known cellulases, but also contain all conserved amino acids identified in our analysis, it is very likely that they also possess cellulolytic activities. On the other hand, GH48s from insects, where only chitinolytic activities were detected experimentally, are non-cellulases. Consequently, the existing Pfam model for GH48 can be used to retrieve true cellulases, however there is one exception: GH48 proteins from insects should be annotated as non-cellulases. This approach allowed us to identify 166 true cellulases in the combined metagenomic dataset of hundreds of environmental samples. The largest number of cellulases came from the metagenomes of 'engineered' microbial communities, such as enriched samples or bioreactors (for example, the 'mixed alcohol bioreactor', and the 'cellulolytic enrichment from sediment of Great Boiling Springs'). Most of the environmental cellulases come from communities that typically include saprophytes (46), such as soil, wastewater, ant fungal gardens and the rhizosphere (Fig. 5), which is in agreement with previously published research (47,48). Interestingly, very few GH48 cellulases were identified in cow rumen microbial communities, which also correlates with previous extensive biochemical analysis of this classical cellulolytic community (29). Moreover, all the GH48s from cow rumen, found in this study, belong to *Ruminococcus flavefaciens*, a highly specialized cellulose-degrader. We hypothesize that because collectively major ruminal cellulolytic specialists are found to represent as little as 0.3% of the total bacterial population (49), and *R. flavefaciens* is typically one of the three most abundant cellulolytic bacteria in cow rumen (50), its GH48 gene was more selective for sequencing (51) when compared to the genes of other 'rare' members of the community.

CONCLUSIONS

High-throughput computational screening for cellulases from genomic and metagenomic datasets is a challenge due to absence of a clear understanding of structural and functional features that distinguish them from closely related enzymes with other substrate specificities (2). Here, we present a combined sequence-structure approach leading to the identification of clear markers that can be used to distinguish between cellulases and non-cellulases within the GH48 family. This approach was applied to identify "true" GH48 cellulases in large metagenomic datasets, illustrating its feasibility in search for novel cellulolytic capabilities.

Finally, we propose that this approach can be generalized to define genomic signatures for identifying cellulases in other CAZy families (2), such as GH5, GH9, GH12, GH45, and GH61, that are known to contain biochemically confirmed cellulases.

REFERENCES

1. Aden, A., and Foust, T. (2009) Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. *Cellulose* **16**, 535-545

2. Sukharnikov, L. O., Cantwell, B. J., Podar, M., and Zhulin, I. B. (2011) Cellulases: ambiguous nonhomologous enzymes in a genomic perspective. *Trends Biotechnol.* **29**, 473-479

3. Cantarel, B. L., and *et al.* (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37 (Database issue)**, D233-238

4. Olson, D. G., Tripathi, S. A., Giannone, R. J., Lo, J., Caiazza, N. C., Hogsett, D. A., Hettich, R. L., Guss, A. M., Dubrovsky, G., and Lynd, L. R. (2010) Deletion of the Cel48S cellulase from *Clostridium thermocellum*. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17727-17732

5. Devillard, E., Goodheart, D. B., Karnati, S. K., Bayer, E. A., Lamed, R., Miron, J., Nelson, K. E., and Morrison, M. (2004) *Ruminococcus albus* 8 mutants defective in cellulose degradation are deficient in two processive endocellulases, Cel48A and Cel9B, both of which possess a novel modular architecture. *J. Bacteriol.* **186**, 136-145

6. Izquierdo, J. A., Sizova, M. V., and Lynd, L. R. (2010) Diversity of bacteria and glycosyl hydrolase family 48 genes in cellulolytic consortia enriched from thermophilic biocompost. *Appl. Eviron. Microbiol.* **76**, 3545-3553

7. Gold, N. D., and Martin, V. J. (2007) Global view of the *Clostridium thermocellum* cellulosome revealed by quantitative proteomic analysis. *J. Bacteriol.* **189**, 6787-6795

8. Wisniewski-Dyé, F., Borziak, K., Khalsa-Moyers, G., Alexandre, G., Sukharnikov, L. O., Wuichet, K., Hurst, G. B., McDonald, W. H., Robertson, J. S., Barbe, V., Calteau, A., Rouy, Z., Mangenot, S., Prigent-Combaret, C., Normand, P., Boyer, M., Siguier, P., Dessaux, Y., Elmerich, C., Condemine, G., Krishnen, G., Kennedy, I., Paterson, A. H., González, V., Mavingui, P., and Zhulin, I. B. (2011) *Azospirillum* genomes reveal transition of bacteria from aquatic to terrestrial environments. *PLoS Genet.* **7**, e1002430

9. Dam, P., Kataeva, I., Yang, S. J., Zhou, F., Yin, Y., Chou, W., Poole, F. L. n., Westpheling, J., Hettich, R., Giannone, R., Lewis, D. L., Kelly, R., Gilbert, H. J., Henrissat, B., Xu, Y., and Adams, M. W. (2011) Insights into plant biomass conversion from the genome of the anaerobic thermophilic bacterium *Caldicellulosiruptor bescii* DSM 6725. *Nucleic Acids Res.* **39**, 3240-3254

10. Irwin, D. C., Zhang, S., and Wilson, D. B. (2000) Cloning, expression and characterization of a family 48 exocellulase, Cel48A, from *Thermobifida fusca*. *Eur. J. Biochem.* **267**, 4988-4997

11. Vazana, Y., Moraïs, S., Barak, Y., Lamed, R., and Bayer, E. A. (2010) Interplay between *Clostridium thermocellum* family 48 and family 9 cellulases in cellulosomal versus noncellulosomal states. *Appl Environ. Microbiol.* **76**, 3236-3243

12. Shen, H., Gilkes, N. R., Kilburn, D. G., Miller, R. C., Jr, and Warren, R. A. (1995) Cellobiohydrolase B, a second exo-cellobiohydrolase from the cellulolytic bacterium *Cellulomonas fimi*. *Biochem. J.* **311**, 67-74

13. Liu, C.-C., and Doi, R. H. (1998) Properties of exgS, a gene for a major subunit of the *Clostridium cellulovorans* cellulosome. *Gene* **211**, 39-47

14. Berger, E., Zhang, D., Zverlov, V. V., and Schwarz, W. H. (2007) Two noncellulosomal cellulases of Clostridium thermocellum, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically. *FEMS Microbiol. Lett.* **268**, 194-201

15. Fujita, K., Shimomura, K., Yamamoto, K., Yamashita, T., and Suzuki, K. (2006) A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. *Biochem. Biophys. Res. Commun.* **345**, 502-507

16. Ramírez-Ramírez, N., Romero-García, E. R., Calderón, V. C., Avitia, C. I., Téllez-Valencia, A., and Pedraza-Reyes, M. (2008) Expression, characterization and synergistic interactions of *Myxobacter sp*. AL-1 Cel9 and Cel48 glycosyl hydrolases. *Int. J. Mol. Sci.* **9**, 247-257

17. Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-W37

18. Katoh, K., and Toh, H. (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899-1900

19. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731-2739

20. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191

21. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402

22. Guindon, S., and Gascuel, O. (2003) PhyML : A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704

23. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Res.* **38** (**Database Issue)**, D211-222

24. Guimarães, B. G., Souchon, H., Lytle, B. L., David Wu, J. H., and Alzari, P. M. (2002) The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the *Clostridium thermocellum* cellulosome. *J. Mol. Biol.* **320**, 587-596

25. Parsiegla, G., Reverbel-Leroy, C., Tardif, C., Belaich, J. P., Driguez, H., and Haser, R. (2000) Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action. *Biochemistry* **39**, 11238-11246

26. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242

27. Koonin, E. V., Makarova, K. S., and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Ann. Rev. Microbiol.* **55**, 709-742

28. Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N., and Kyrpides, N. C. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40 (Database issue)**, D115-122

29. Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennacchio, L. A., Tringe, S. G., Visel, A., Woyke, T., Wang, Z., and Rubin, E. M. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463-467

30. Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., and Crocker, D. (2006) LAP - determination of structural carbohydrates and lignin in biomass. *Technical Report* NREL/TP-510-42618.

31. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton,

E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235-242

32. Vagin, A., and Teplyakov, A. (2010) Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 22-25

33. Langer, G., Cohen, S. X., Lamzin, V. S., and Perrakis, A. (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171-1179

34. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486-501

35. Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355-367

36. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12-21

37. Engh, R. A., and Huber, R. (1991) Accurate Bond and Angle Parameters for X-Ray Protein-Structure Refinement. *Acta Crystallogr. A* **47**, 392-400

38. Gutiérrez-Preciado, A., Henkin, T. M., Grundy, F. J., Yanofsky, C., and Merino, E. (2009) Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiol. Mol. Biol. Rev.* **73**, 36-61

39. Koonin, E. V. (2005) Orthologs, paralogs and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309-338

40. Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256-2268

41. Parsiegla, G., Juy, M., Reverbel-Leroy, C., Tardif, C., Belaïch, J. P., Driguez, H., and Haser, R. (1998) The crystal structure of the processive endocellulase CelF of *Clostridium cellulolyticum* in complex with a thiooligosaccharide inhibitor at 2.0 A resolution. *EMBO J.* **17**, 5551-5562

42. Parsiegla, G., Reverbel, C., Tardif, C., Driguez, H., and Haser, R. (2008) Structures of mutants of cellulase Cel48F of *Clostridium cellulolyticum* in complex with long hemithiocellooligosaccharides give rise to a new view of the substrate pathway during processive action. *J. Mol. Biol.* **375**, 499-510

43. Pignatelli, M., and Moya, A. (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One* **6**, e19984

44. Rho, M., Tang, H., and Ye, Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191

45. Fetrow, J. S. (1995) Omega loops: nonregular secondary structures significant in protein function and stability. *The FASEB Journal* **9**, 708-717

46. Medie, F. M., Davies, G. J., Drancourt, M., and Henrissat, B. (2012) Genome analyses highlight the different biological roles of cellulases. *Nat. Rev. Microbiol.* **10**, 227-234

47. Suen, G., and et al. (2010) An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genet.* **6,** e1001129

48. Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., Mitter, B., Hauberg-Lotte, L., Friedrich, F., Rahalkar, M., Hurek, T., Sarkar, A., Bodrossy, L., van Overbeek, L., Brar, D., van Elsas, J. D., and Reinhold-Hurek, B. (2012) Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Mol. Plant Microbe Interact.* **25**, 28-36

49. Brulc, J. M., Yeoman, C. J., Wilson, M. K., Berg Miller, M. E., Jeraldo, P., Jindou, S., Goldenfeld, N., Flint, H. J., Lamed, R., Borovok, I., Vodovnik, M., Nelson, K. E., Bayer, E. A., and White, B. A. (2011) Cellulosomics, a gene-centric approach to investigating the intraspecific diversity and adaptation of *Ruminococcus flavefaciens* within the rumen. *PLoS One* **6**, e25329

50. Huws, S. A., Lee, M. R., Muetzel, S. M., Scott, M. B., Wallace, R. J., and Scollan, N. D. (2010) Forage type and fish oil cause shifts in rumen bacterial diversity. *FEMS Microbiol Ecol.* **73**, 396-407

51. Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., and Wittwer, P. (2005) Metagenomic gene discovery: past, present and future. *Trends Biotechnol.* **23**, 321-329

ACKNOWLEDGEMENTS

This work was supported by the DOE Office of Science, Office of Biological and Environmental Research, through the BioEnergy Science Center (BESC), a DOE Bioenergy Research Center.

Table legends:

Table 1. Enrichment of GH48 genes in the prokaryotic genomes.

Figure legends:

Figure 1. Modular domain architecture of GH48 paralogs. Representative examples of GH48 paralogs that contain different auxiliary domains are shown. The GenBank identifiers (GI numbers) are listed next to each protein. Protein domains: CBM, carbohydrate-binding module; Dock, dockerin; DUF, domain of unknown function; BNR, BNR repeat; GH, glycoside hydrolase.

Figure 2. Horizontal gene transfer of GH48 enzymes. A maximum-likelihood phylogenetic tree constructed from multiple sequence alignment of GH48 sequences is shown. Known enzymatic activities, taxonomic information and inferred evolutionary relationships are shown on the outside circle. Sequences from underrepresented phyla are marked with an asterisk: (1) Proteobacteria and (2) Chloroflexi.

Figure 3. Structure of GH48 from *Hahella chejunsis*. Additional omega loop identified in all cellulases is labeled in blue. The α-helices are shown in red, β-strands in yellow, loops in green and the cellobiose molecule with carbon atoms in cyan and oxygen atoms in red.

Figure 4. Phyletic distribution of GH48 sequences retrieved from a combined metagenomic dataset. Nearly 95% of sequences belong to three closely related prokaryotic phyla: Actinobacteria, Firmicutes and Chloroflexi.

Figure 5. Abundance of GH48 cellulases in metagenomes. A. Percentage of GH48 sequences in each metagenome (abundance) was calculated by dividing the number of GH48 hits by the total number of genes in each metagenome. B. The abundance of GH48 sequences in different habitats. The normalized percentage of GH48 genes was calculated as percentage of GH48 sequences in a given metagenome divided by the sum of percentage of GH48 for all metagenomes.

Table 1

Taxon	Total number of genomes*/Number of genomes containing GH48 genes	Percentage of genomes containing GH48 genes
Actinobacteria	218/38	17
Firmicutes	418/80	19
Chloroflexi	17/2	12
Proteobacteria	732/2	0.3

*Complete and draft genomes with a size >1 Mb

Figure 1

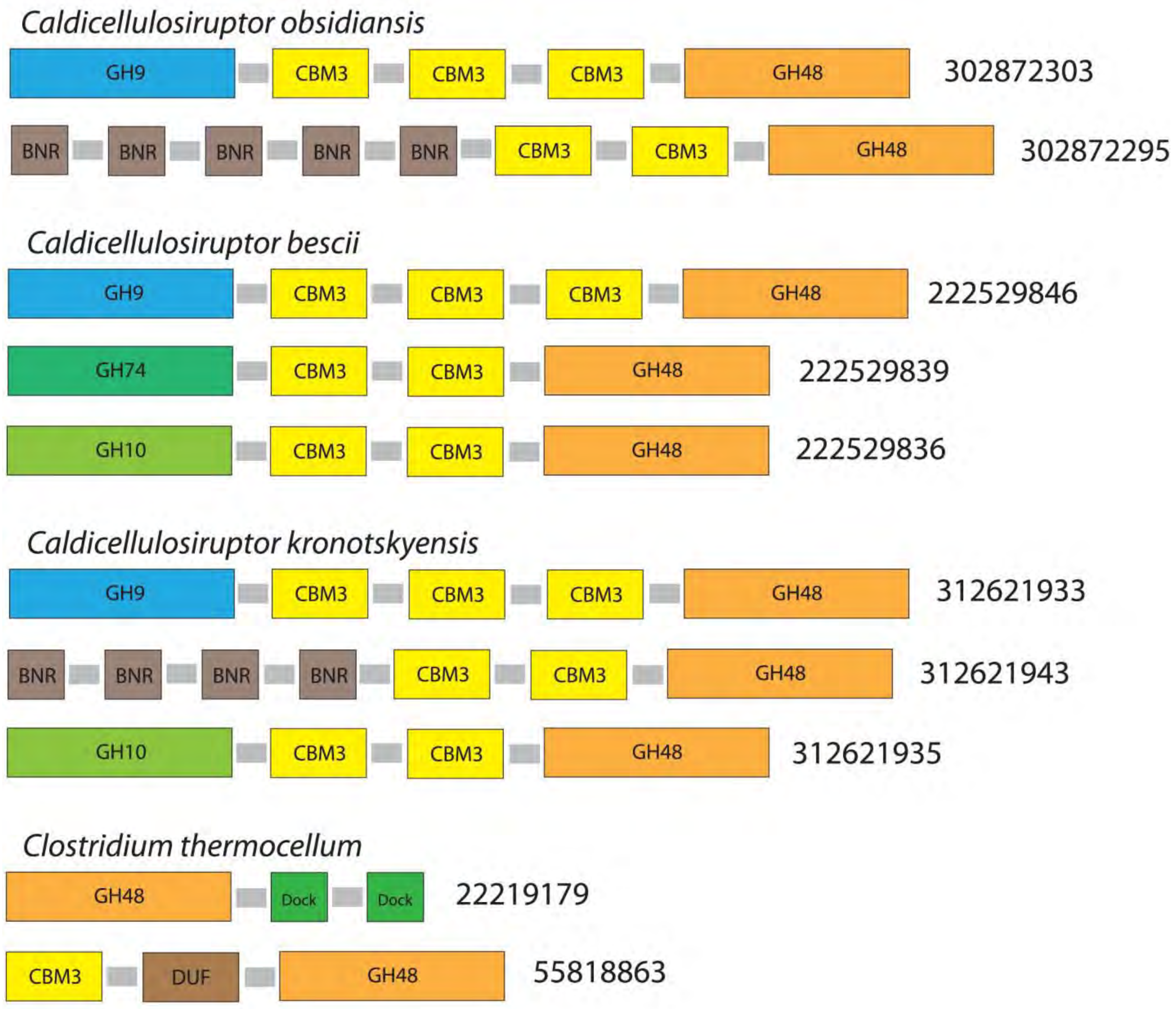

Caldicellulosiruptor obsidiansis
GH9
CBM3
CBM3
CBM3
GH48
302872303
BNR
BNR
BNR
BNR
BNR
CBM3
CBM3
GH48
302872295
Caldicellulosiruptor bescii
GH9
CBM3
CBM3
CBM3
GH48
222529846
GH74
CBM3
CBM3
GH48
222529839
GH10
CBM3
CBM3
GH48
222529836
Caldicellulosiruptor kronotskyensis
GH9
CBM3
CBM3
CBM3
GH48
312621933
BNR
BNR
BNR
BNR
CBM3
CBM3
GH48
312621943
GH10
CBM3
CBM3
GH48
312621935
Clostridium thermocellum
GH48
Dock
Dock
22219179
CBM3
DUF
GH48
55818863

Figure 2

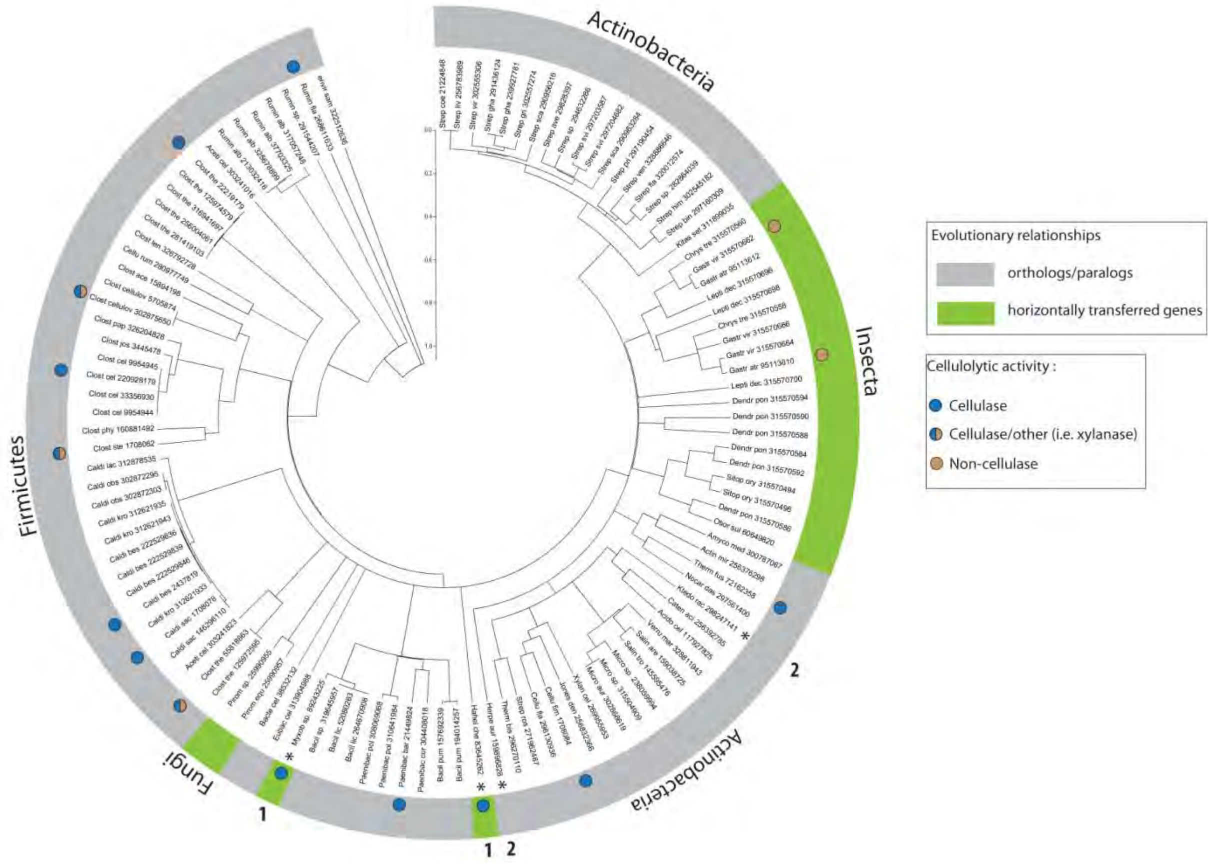

Actinobacteria
Insecta
Actinobacteria
Firmicutes
Fungi
1
1 2
2
*
Evolutionary relationships
orthologs/paralogs
horizontally transferred genes
Cellulolytic activity :
Cellulase
Cellulase/other (i.e. xylanase)
Non-cellulase

Figure 3

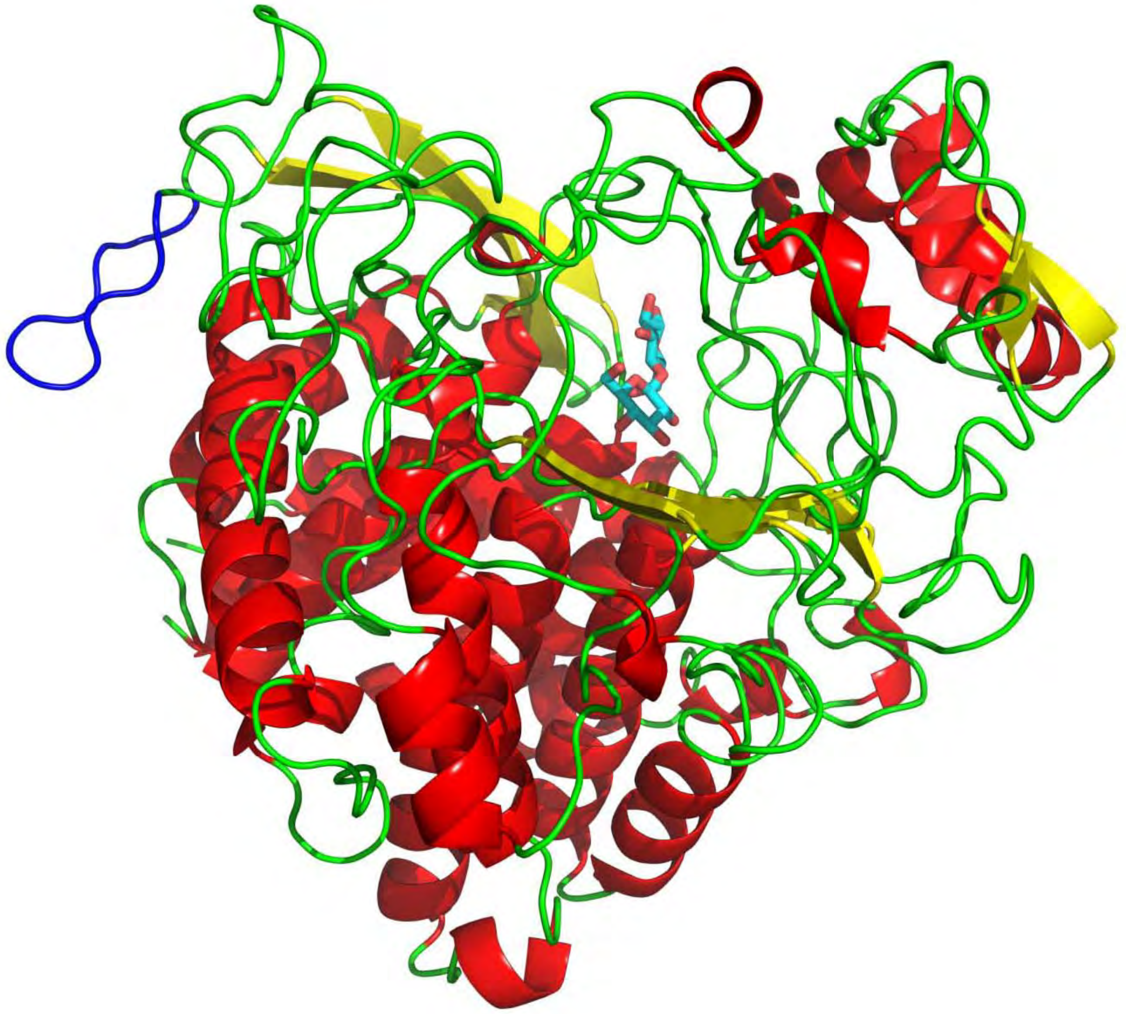


Figure 4

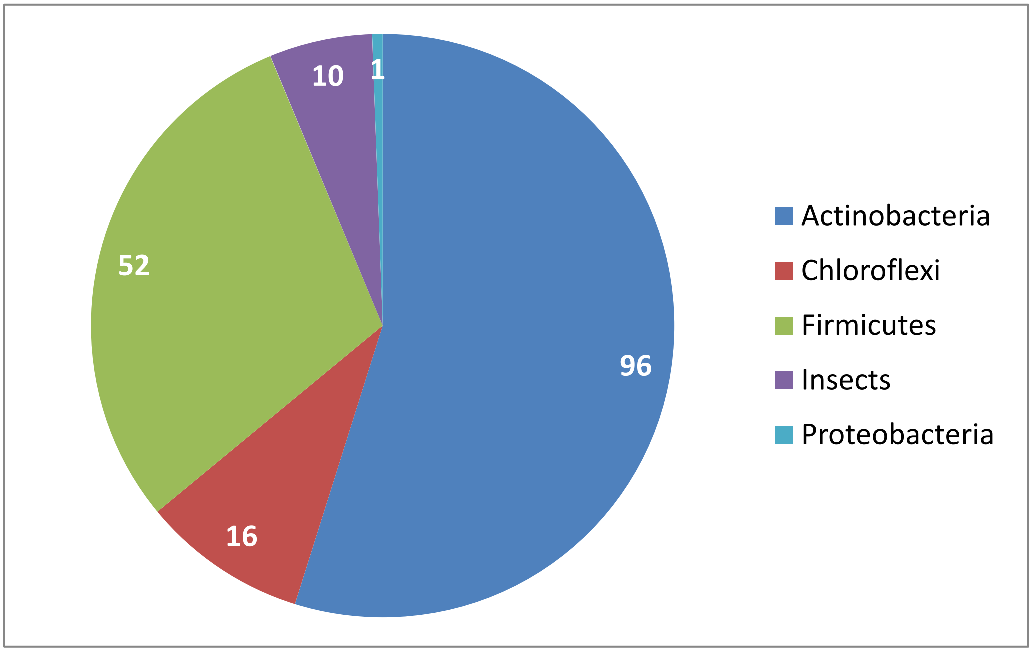
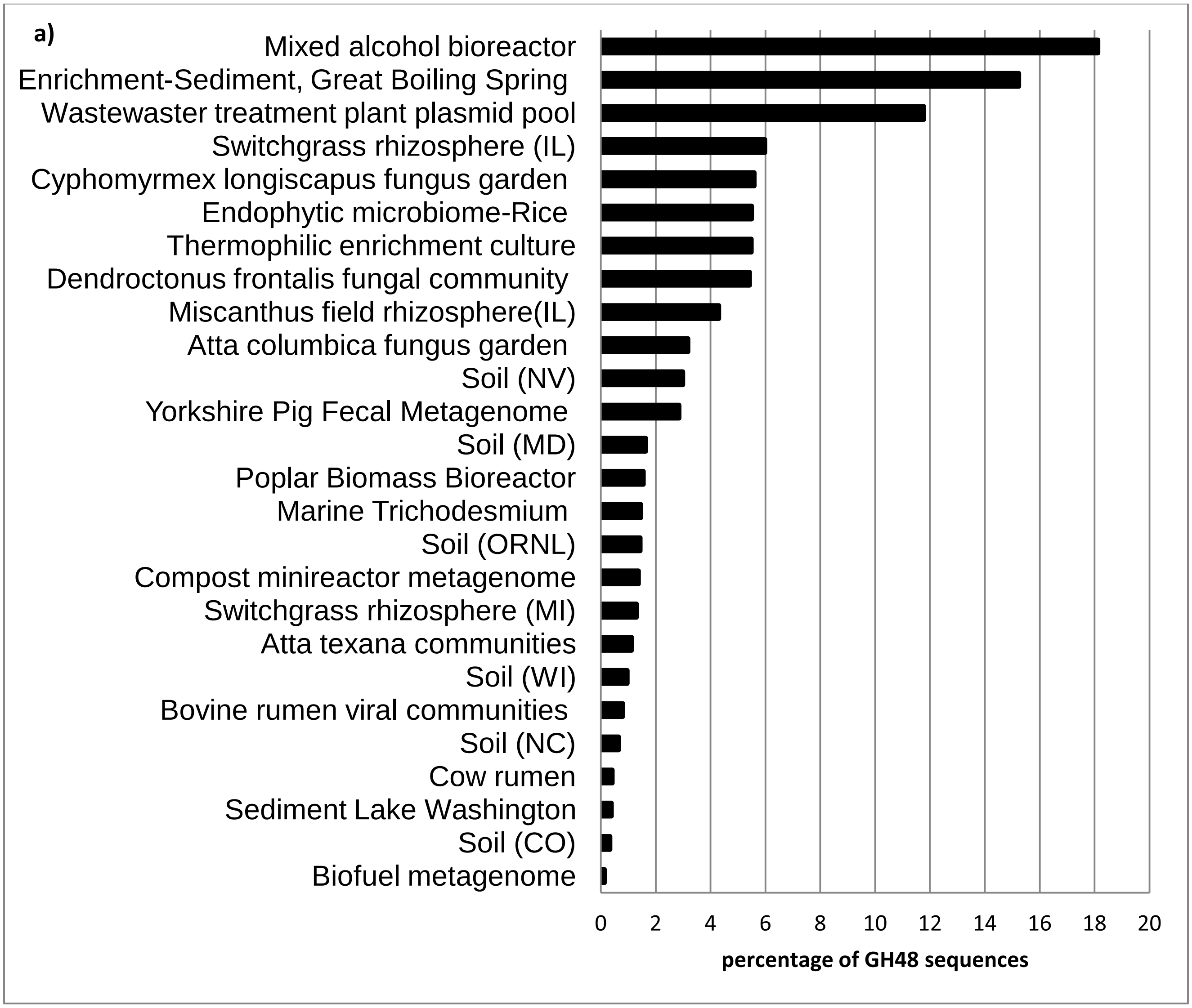
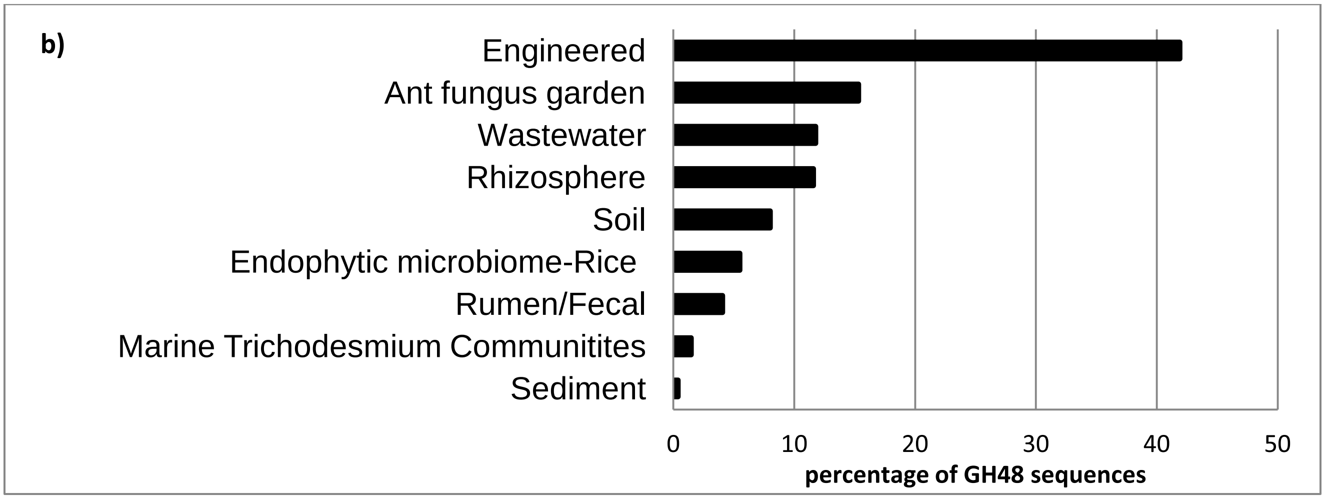

10
1
52
96
16
Actinobacteria
Chloroflexi
Firmicutes
Insects
Proteobacteria

Figure 5


a)
Mixed alcohol bioreactor
Enrichment-Sediment, Great Boiling Spring
Wastewaster treatment plant plasmid pool
Switchgrass rhizosphere (IL)
Cyphomyrmex longiscapus fungus garden
Endophytic microbiome-Rice
Thermophilic enrichment culture
Dendroctonus frontalis fungal community
Miscanthus field rhizosphere(IL)
Atta columbica fungus garden
Soil (NV)
Yorkshire Pig Fecal Metagenome
Soil (MD)
Poplar Biomass Bioreactor
Marine Trichodesmium
Soil (ORNL)
Compost minireactor metagenome
Switchgrass rhizosphere (MI)
Atta texana communities
Soil (WI)
Bovine rumen viral communities
Soil (NC)
Cow rumen
Sediment Lake Washington
Soil (CO)
Biofuel metagenome
0
2
4
6
8
10
12
14
16
18
20
percentage of GH48 sequences


b)
Engineered
Ant fungus garden
Wastewater
Rhizosphere
Soil
Endophytic microbiome-Rice
Rumen/Fecal
Marine Trichodesmium Communitites
Sediment
0
10
20
30
40
50
percentage of GH48 sequences