AEM
Journals.ASM.org

Phylogenetic Distribution of Potential Cellulases in Bacteria

Renaud Berlemont, Adam C. Martiny

Department of Earth System Science, and Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, California, USA

Many microorganisms contain cellulases that are important for plant cell wall degradation and overall soil ecosystem functioning. At present, we have extensive biochemical knowledge of cellulases but little is known about the phylogenetic distribution of these enzymes. To address this, we analyzed the distribution of 21,985 genes encoding proteins related to cellulose utilization in 5,123 sequenced bacterial genomes. First, we identified the distribution of glycoside hydrolases involved in cellulose utilization and synthesis at different taxonomic levels, from the phylum to the strain. Cellulose degradation/utilization capabilities appeared in nearly all major groups and resulted in strains displaying various enzyme gene combinations. Potential cellulose degraders, having both cellulases and β-glucosidases, constituted 24% of all genomes whereas potential opportunistic strains, having β-glucosidases only, accounted for 56%. Finally, 20% of the bacteria have no relevant enzymes and do not rely on cellulose utilization. The latter group was primarily connected to specific bacterial lifestyles like autotrophy and parasitism. Cellulose degraders, as well as opportunists, have multiple enzymes with similar functions. However, the potential degraders systematically harbor about twice more β-glucosidases than their potential opportunistic relatives. Although scattered, the distribution of functional types, in bacterial lineages, is not random but mostly follows a Brownian motion evolution model. Degraders form clusters of relatives at the species level, whereas opportunists are clustered at the genus level. This information can form a mechanistic basis for the linking of changes in microbial community composition to soil ecosystem processes.

The degradation of plant cell wall polymers by soil microorganisms is a key step for carbon cycling in many ecosystems. Bacteria and fungi express enzymes for the decomposition of plant cell wall polymers (1). However, not all soil microorganisms carry this ability (2), and some soil bacteria—the opportunists—benefit from the activity of enzymes produced by other organisms—the degraders (3, 4). Thus, the phylogenetic distribution of enzyme genes involved in plant cell degradation is likely important for understanding the link between microbial community composition and soil ecosystem processes (2, 5). Recent theoretical models for litter decay based on functional trait distribution (6) or functional subpopulations (guilds) (7) provide new insights into the interactions of the environmental variables (e.g., substrate availability) and the dynamics of microbial populations. These approaches require precise trait-specific phylogenomic frameworks in order to connect the functional traits or populations to phylogenetically described bacterial lineages. But at present, we have little knowledge of the overall phylogenetic distribution of many enzymes involved in plant-derived polymers.

The superfamily of glycoside hydrolases (GH) is an important group of enzymes for plant cell wall degradation (8). GH are active on glycosidic bonds between carbohydrates or between carbohydrates and noncarbohydrate moieties. They include enzymes having different folds and catalyzing many reactions (9). Some GH families catalyze a specific reaction, whereas other activities are found in multiple families (Table 1). Cellulose degradation is regarded as a complex process requiring the synergistic action of multiple GH families (8). These families include at least three described types of proteins active on β-1,4 glycosidic bonds: (i) endocellulases, active on internal β-1,4 glucosidic bonds, (ii) exocellulases degrading the polymer by its extremities, and (iii) β-glucosidases producing glucose from cellobiose (Table 1) (9). Finally, many GH, including cellulases, are frequently isolated as modular enzymes associated with other catalytic and noncatalytic domains like carbohydrate binding modules (CBM) (10, 11). Once generated outside the cell, cellooligosaccharides produced by the GH enter the cytoplasm using "cellobiose-specific" uptake systems such as the phosphotransferase system (12) or using ABC transporters (13). Otherwise, glucose has to be generated outside the cell by secreted β-glucosidases and imported to the cytoplasm by "glucose transporters" (14). Inside the cell, phosphorylated sugars are channeled to the central metabolism.

The biochemistry (the catalytic mechanisms, the kinetics, the substrate binding) of many GH is well characterized (1, 15–18). This information, together with comparisons of some structures and sequences, resulted in an accurate GH classification (9, 19, 20). Moreover, a few cellulolytic strains have also been subjected to thorough characterization and to sequencing of their genomes (1, 15). Recently, GH-specific annotation systems were introduced in order to precisely assign function to newly sequenced GH (21, 22). Nevertheless, little is known about the global phylogenetic distribution of the cellulose utilization potential in microorganisms. Previous works have suggested that the ability to grow on many carbon sources is generally not randomly distributed in bacteria. Instead, it is shared by bacteria forming a cluster of closely related taxa (23, 24). This raises the following questions: (i) what is the phylogenetic distribution of genes for cellulose utilization and (ii) are there differential phylogenetic distributions and clusterings of enzymes involved in the different steps of cellulose degradation/cellobiose utilization? To answer these questions, we

Received 30 October 2012 **Accepted** 17 December 2012

Published ahead of print 21 December 2012

Address correspondence to Adam C. Martiny, amartiny@uci.edu.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.03305-12.

Copyright © 2013, American Society for Microbiology. All Rights Reserved.

doi:10.1128/AEM.03305-12

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

TABLE 1 Abundance of bacterial sequences from glycoside hydrolase families associated with the degradation of carbon molecules with β-glucosidic bonds[a]

GH family	Genes described in the CAZy database						Cellulose synthesis (reference)	No. of identified orthologs (SEED)
	Total no. of hits	Characterized enzymes[b]	Endocellulase	Exocellulase	β-Glucosidase	Other characterized activity (hits)		
1	2,890	87		1	48	5 (43)		10,364
3	3,212	91		5	55	8 (50)		8,739
5	1,952	206	152	2		6 (63)	Yes (39)	1,368
6	231	16	13	4			Yes (40)	140
8	538	55	26			4 (33)	Yes (31)	340 GH8 + 616 BcsZ
9	468	75	66	6	1	1 (1)		231
12	144	22	21			1 (1)		126
44	55	10	8			1 (4)		3
45	13	13	3					6
48	128	13	4	11				52

[a] Data include genes described in the CAZy database (9), the number of characterized bacterial enzymes, and the presence of these genes in all bacterial sequenced genomes in the SEED database (as of 1 March 2012) (27). Known displayed activity data represent such activities in characterized enzymes as described in CAZy. Cellulose synthesis data represent the involvement of enzymes in cellulose synthesis. Identified ortholog data represent the number of genes identified in the present work.
[b] Some characterized proteins display several activities.

present a new approach to connect the distribution of functional traits to the phylogeny of known bacterial lineages. Our results can provide a mechanistic basis for determining whether or not we should expect a link between microbial community composition and the cellulose degradation, a major prerequisite for ecosystem modeling.

MATERIALS AND METHODS

Identification of genes encoding glycoside hydrolases. Families of orthologous genes for cellulose degradation were identified following the CAZy database classification scheme (http://www.cazy.org/) (9). For each GH family containing enzymes involved in cellulose utilization (e.g., GH5), the corresponding protein sequences were retrieved from the PFam database (e.g., PF00150) (25). For all these proteins, the corresponding FIGfam numbers (26) were then extracted from the SEED database (http://www.theseed.org) using the SEED API server (e.g., for GH5, FIG 00002614 to FIG 00003086, among others) (27). The SEED annotation system was considered a robust annotation system for the analysis of sequenced genomes (28). The functions associated with the selected FIGfams were manually checked. Finally, FIGfam numbers were used to mine the FIGfam-annotated genomes hosted in the genome database of the Pathosystems Resource Integration Center (PATRIC, http://www.patricbrc.org) (29). Retrieved protein sequences were scanned against the PFam database using PFam-scan in order to check the annotation (25). Finally, the number of occurrences of proteins from all the specified GH families was compiled for further analysis (see Table S1 and Fig. S1 in the supplemental material).

Following the CAZy database, ~6% of the proteins from the relevant GH families had been characterized. The functional annotation of the uncharacterized genes resulted from stepwise annotation transfer from similar characterized or uncharacterized enzymes. However, even closely related and characterized proteins could have different substrate specificities. Thus, most of the annotations for uncharacterized enzymes were still considered hypothetical.

The same approach, starting from the PFam database, was used to identify the distribution of the RubisCO large subunit and some genetic markers associated with pathogenic strains in the sequenced bacterial genomes (30). Glycoside hydrolases from the family GH8 and involved in the production of cellulose in bacteria (*bcsZ*) were identified as they were located in the conserved "bacterial cellulose synthesis" (BCS) operon (31).

Phylogenetic analysis. Aligned 16S rRNA sequences corresponding to the 5,123 analyzed genomes were extracted from the SILVA database (http://www.arb-silva.de/) (32). Next, a 16S rRNA phylogenetic tree was constructed using PHYLIP (neighbor joining, distance matrix F84, 100 bootstraps) (33, 34) and displayed using the ITOL server (35).

GH similarity profile clustering and comparisons of matrices. The strain-specific GH profiles for the 5,123 considered bacteria were compared using the Bray-Curtis similarity index. The similarity matrix produced was used for hierarchical clustering, using the Vegan package in R software (36). Correlations between the phylogenetic distance matrices (generated using PHYLIP) and the GH-based similarity matrices were computed by applying a Mantel test using the R package for Analysis of Phylogenetics and Evolution (APE) (37).

Phylogenetic conservation and clustering. The phylogenetic signal strength in binary traits (D) was tested running the Fritz and Purvis test with 1,000 permutations (38), using the R package "CAPER." A D value below 0 reveals a strongly clumped distribution, $D \cong 0$ means a "Brownian motion"-like evolutionary distribution, $D \cong 1$ a random distribution, and $D > 1$ an overdispersed distribution. The "consenTRAIT" algorithm for trait depth analysis (23) was used to evaluate the phylogenetic conservation of genotypes. consenTRAIT determines the size of clusters of strains containing at least 90% of relatives sharing a common trait. The trait depth (τ_D) of the cluster was then computed as the consensus 16S rRNA distance from the tips to the node corresponding to the last common ancestor of the cluster.

RESULTS

To identify the phylogenetic distribution of traits involved in the cellulose utilization, we first extracted a list of genes belonging to glycoside hydrolases (GH) for cellulose deconstruction using the CAZy database (Table 1). We then linked these genes to orthologous gene families in the SEED database (26). This allowed us to identify putative GH families in 5,123 sequenced bacterial genomes (see Table S1 in the supplemental material). Enzymes from the families GH1 and GH3 were the most abundant enzymes detected in the bacterial genomes (Table 1). Following the CAZy database, 55% and 60% of enzymes characterized from these two families appeared to be β-glucosidases. The non-β-glucosidases displayed several activities, including exocellulase (1% and 5%), 6-phospho-β-glucosidase (29% and 0%), and a few other activities. Thus, the majority of the genes from families GH1 and GH3 were related to the processing of small oligo- and disaccharides. GH5 was the most abundant protein family almost exclusively displaying endocellulase activity (74% of the characterized proteins), although a few (~1%) GH5 proteins were also classified as

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

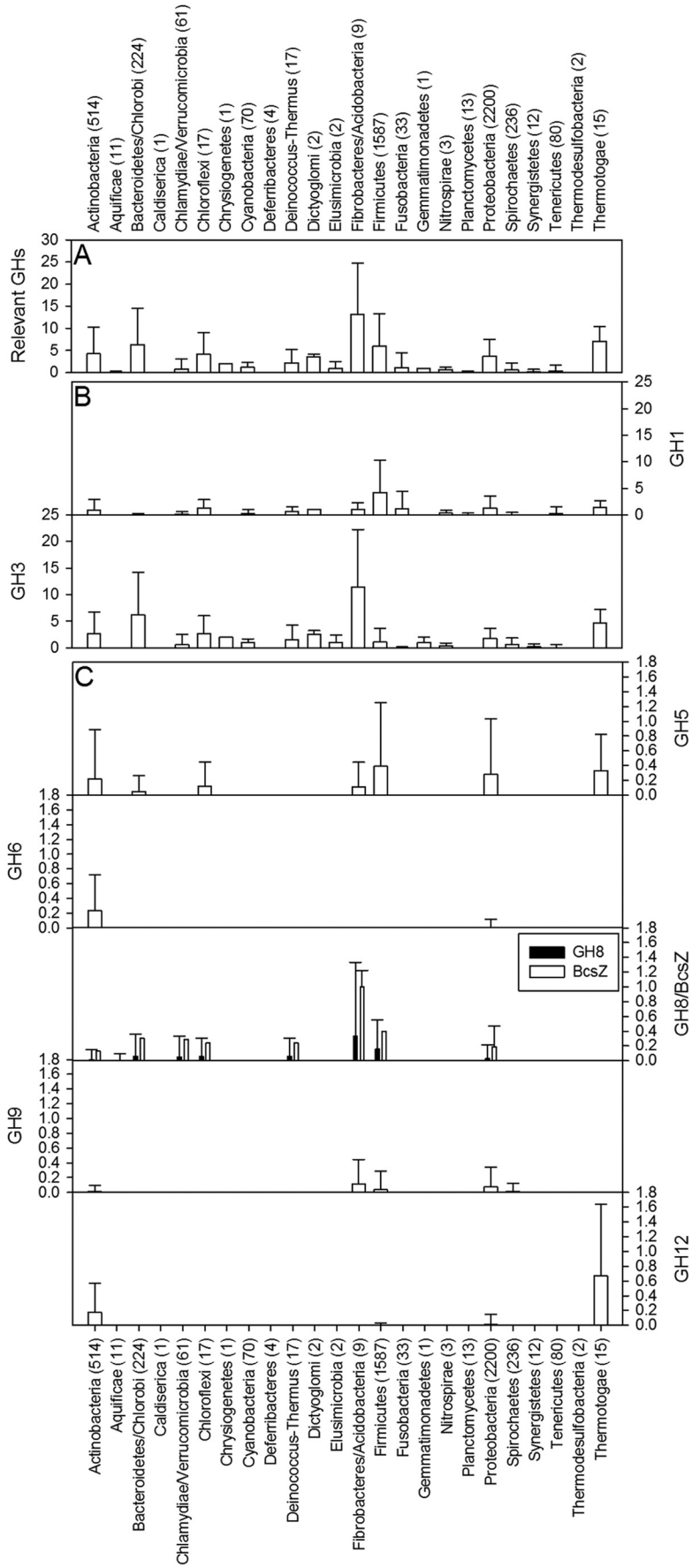


FIG 1 Phylum-specific average glycoside hydrolase gene content (expressed as genes/genome). The number of analyzed genomes from each phylum is indicated in brackets beside the phylum names.

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

exocellulases. The noncellulolytic GH5 included β-mannanase activity as well as licheninase or chitosanase. With few exceptions, enzymes from the GH8 family were endocellulases (48%) or chitosanases (40%). GH6 and GH9 were more specific for complex cellulose polymer degradation, as 81% and 88% of the characterized enzymes were endocellulases, whereas 25% and 9% displayed exocellulase activity, respectively. Notably, some of these enzymes had both activities. The enzymes from family GH12 were nearly all endocellulases (95%). Enzymes from families GH44 and GH45 were generally endo- or exocellulases, whereas those from family GH48 were almost exclusively exocellulases. Proteins from families GH44, -45, and -48 are frequently attached to other domains (catalytic and/or carbohydrate binding module) and sometimes part of complex cellulosomes. They were exclusively found in cellulolytic *Clostridia*, *Fibrobacteres*, and *Paenibacillus* lineages and can therefore be regarded as highly specialized enzyme families. Overall, as described in CAZy (9), it appeared that β-glucosidases responsible for the processing of small oligosaccharides were predominantly found in GH1 and GH3, whereas the endo- and exocellulases were mostly present in GH5, -6, -8, -9, -12, -44, -45, and -48.

Distribution of GH in bacterial phyla. The abundance of genes encoding putative GH for cellulose utilization varied across bacterial phyla, but these genes were not detected in members of several phyla, including *Caldiserica*, *Deferribacteres*, and *Thermodesulfobacteria* (Fig. 1A). In addition, many phyla possessed less than one GH/genome. In contrast, 10 phyla contained more than one GH, for cellulose utilization, per genome. Finally, members of several phyla, including the *Firmicutes*, *Bacteroidetes*/*Chlorobi* group, *Fibrobacteres*, *Acidobacteria*, and *Thermotogae*, on average possessed more than 5 putative enzymes per genome.

Across all phyla, GH1 and GH3 were the most abundant families (Fig. 1B), and 79% of the analyzed genomes contained a least one gene from one of the two families. Genes from these families were found in nearly all the phyla—with a few exceptions (*Caldiserica*, *Aquificae*, *Deferribacteres*, and *Thermodesulfobacteria*). Genes associated with GH5 were the most abundant putative endo- and exocellulase-encoding genes and were detected in the *Actinobacteria*, *Bacteroidetes*/*Chlorobi* group, *Chloroflexi*, *Fibrobacteres*/*Acidobacteria*, *Firmicutes*, *Proteobacteria*, and *Thermotogae* (Fig. 1C). Other GH families specific for the polymer cellulose hydrolysis (i.e., GH6, -8, -9, -12, -44, -45, and -48) were less abundant in the bacterial genomes (Fig. 1C; see also Fig. S2 in the supplemental material). GH6 was solely detected in *Actinobacteria* and *Proteobacteria*. Similarly, genes associated with GH9 were observed only in the genomes of some *Actinobacteria*, *Fibrobacteres*, *Firmicutes*, *Proteobacteria*, and *Spirochaetes*, whereas those associated with GH12 were observed exclusively in *Actinobacteria*, *Thermotogae*, *Firmicutes*, and *Proteobacteria*. Finally, genes associated with GH44, -45, and -48 were found in the genomes of some fully sequenced cellulolytic *Actinobacteria*, *Firmicutes*, and *Proteobacteria*.

We also identified the distribution of putative biosynthetic enzymes from GH8 and found that 64% of genes from this family were likely part of a conserved bacterial cellulose synthesis (BCS) operon (Fig. 1C). In *Proteobacteria*, more than 75% of GH8 enzyme-encoding genes were likely biosynthetic. Although less abundant, these genes were also commonly observed in *Aquificae* and in *Fibrobacteres*/*Acidobacteria*.

Co-occurrence of glycoside hydrolases. We next determined the frequency of genomically encoded "functional types" for cellulose degradation in microorganisms. We predicted that 20% of bacteria did not have genes for GH involved in cellulose degradation and were thus unable to utilize cellulose as a carbon or energy source. A total of 56% of the lineages contained only predicted genes for β-glucosidase (families GH1 and/or GH3). Such strains could be considered potential opportunists, as they may use only cellobiose generated by the degradation of longer polymers by other lineages. A total of 24% of the bacterial genomes possessed genes for hypothetical cellulase and β-glucosidase and could be regarded as potential cellulose degraders. Finally, less than 1% of the analyzed genomes contained a gene(s) for cellulase but not for β-glucosidase. In these genomes, genes encoding GH from families GH5, -6, -8, -9, -12, -44, -45, and -48 may produce an enzyme other than cellulose (e.g., endo-β-1,4-mannosidase from the GH5 family) or cellulases involved in cellulose production (39, 40), nonhydrolytic interaction (41), or endophytic colonization (42) instead of cellulose degradation. These pathways do not require β-glucosidases. Alternatively, we may not have identified all β-glucosidases in these sequenced genomes.

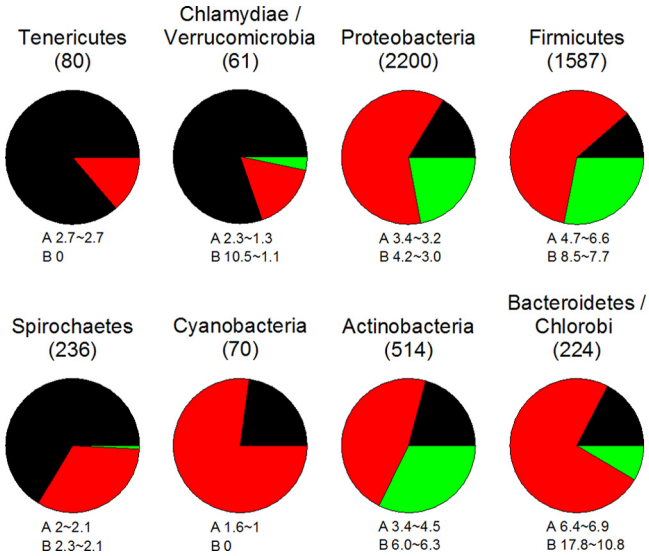


FIG 2 Distribution of cellulose utilization strategies in fully sequenced genomes of the eight most abundant phyla. The fractions include bacteria having no GH relevant for cellulose utilization (black), potential opportunists (red), and potential cellulose degraders (green). A and B values represent the average β-glucosidase content (~standard deviation [SD]) in opportunists versus potential cellulose degraders.

We observed extensive variations of these three functional types ("no interest," potential opportunist, and potential degrader) within each phylum. Potential cellulose degraders accounted for less than 32% of the strains and were rare or absent from several phyla (Fig. 2). However, the potential degraders were common in *Actinobacteria*, *Firmicutes*, *Proteobacteria*, and *Bacteroidetes*/*Chlorobi* and accounted for 32%, 27%, 22%, and 8% of the sequenced strains, respectively. We also found an elevated number of genes for putative β-glucosidases in potential degraders (Fig. 2, B value) compared to potential opportunists (Fig. 2, A value). Considered as a whole, 50% of the potential opportunists had 1 or 2 β-glucosidases, whereas more than 75% of the potential cellulose degraders had more than 2 of these enzymes. Also, 50% of the strains defined as potential cellulose degraders possessed

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

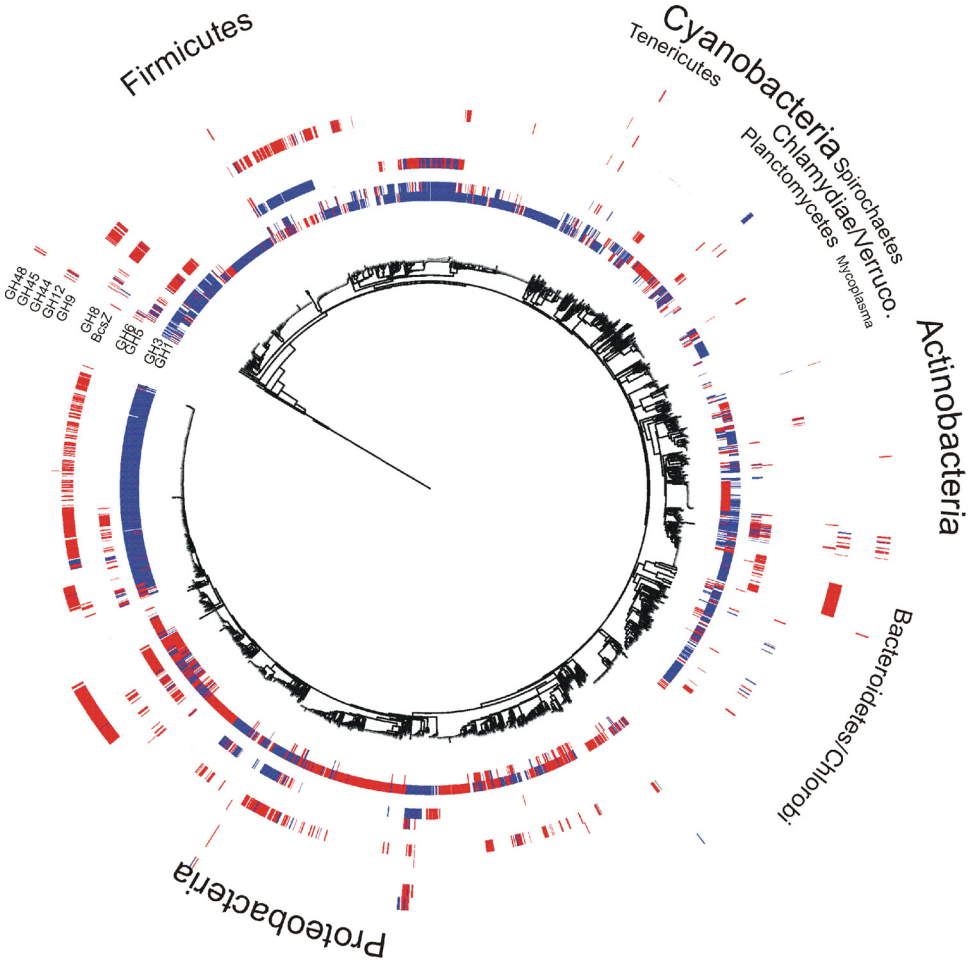


FIG 3 Strain-specific glycoside hydrolase distribution in bacteria. The consensus phylogenetic tree was based on a 16S rRNA alignment retrieved from the SILVA database (32) and constructed using Phylip (distance matrix F84, neighbor joining, 100 bootstraps) (34). The outer circles show the absence (white) or the presence of one copy (red) or multiple copies (blue) of genes from a considered GH family in each genome.

several orthologous genes for possible cellulases (see Fig. S3 in the supplemental material).

In our data set, most of the bacteria appear to be probable opportunists. These organisms have only a predicted β-glucosidase activity—albeit at times they contained multiple genes for this activity. The potential cellulose degraders generally contained several putative cellulases and many β-glucosidases. However, potential opportunists and degraders had GH content that varied extensively even at the intraphylum level. This suggested that some important variations occurred at lower taxonomic levels.

Phylogenetic distribution of GH families at lower taxonomic levels. Since the GH content was not homogeneously distributed within each phylum, we next investigated the detailed phylogenetic distribution of GH. This was done in order to identify clusters of organisms sharing similar GH content and to determine their phylogenetic relatedness. First, we evaluated the overall relationship between the 16S rRNA phylogeny and the global GH distribution (Fig. 3) but observed no clear correlation (Mantel test, r_m = 8.81e-6, P = 0.5). Indeed, Fig. 1 and 3 illustrate the extensive variability in the GH content across the bacterial tree. Within each GH group, we saw a slightly different pattern. β-Glucosidases from GH1 and GH3 were present in most lineages and followed a nonrandom phylogenetic distribution (Table 2). The genes from these two families were generally clustered in clades with a clade depth (τ_D) of ~0.02 16S rRNA distance (Table 2). The GH genes encoding endo- and exocellulases were also nonrandomly phylogenetically dispersed but occurred in cluster sizes that were on average smaller than those seen with GH1 and GH3. However, we mainly detected these cellulase clusters in *Proteobacteria*, *Firmicutes*, and *Actinobacteria*. Some of these clusters included groups of well-characterized cellulolytic organisms such as *Clostridia* and *Streptomyces*.

Next, we examined the phylogenetic distribution of the functional types over our entire data set (Table 2). Overall, we found that the potential degraders formed small clusters (~4.3 bacteria/cluster) of closely related strains (τ_D = 0.013 16S rRNA distance). In contrast, the strains having only genes for GH1 or -3 formed bigger clusters (~12.2 bacteria/cluster). Thus, our analysis showed that GH as well as functional types for cellulose degradation in bacteria were phylogenetically nonrandom and more resembled a Brownian motion-type evolution. The smaller clades of potential cellulose degraders suggested some specialization of these organisms.

We also analyzed the distribution of the different strategies in

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

TABLE 2 Phylogenetic distribution clustering of glycoside hydrolase families and functional types[a]

Category	Trait	D	P_{random}	$P_{Brownian}$	ACS	τ_D
GH family (hydrolase only)	GH1	0.026	0	0.292	5.607	0.020
	GH3	0.079	0	0.049	7.180	0.024
	GH5	−0.015	0	0.609	3.865	0.010
	GH6	−0.016	0	0.578	1.603	0.005
	GH8	0.016	0	0.440	2.008	0.009
	GH9	−0.352	0	0.998	3.772	0.009
	GH12	−0.280	0	0.953	4.268	0.009
	GH44	0.412	0	0.122	1	0.028
	GH45	0.497	0.157	0.303	1	0.033
	GH48	0.409	0	0.097	1.199	0.011
Function	BcsZ	0.135	0	0.033	2.156	0.005
	β-Glucosidase	0.111	0	0.036	12.163	0.034
	Cellulase	−0.006	0	0.566	4.355	0.013

[a] Significance of clustering is based on the Fritz and Purvis index (D) (38) and compared to the probability that the considered trait will follow "random" or "Brownian" evolution. Average cluster size (ACS) and depth (τ_D, in 16S rRNA sequence distance) were estimated using consenTRAIT (23).

specific phylogenetic groups (Fig. 3). We found that the GH content similarity increased when more closely related strains were considered. To illustrate this point, we analyzed the distribution of GH families in detail in *Actinobacteria*, as most of their subgroups have been intensively studied and sequenced (43). These organisms were also regarded as important for cellulose degradation in soil (2). In the 514 analyzed genomes from this phylum, 20% of the genomes contained no gene for GH related to the cellulose utilization, 47% were considered potential opportunists, and 32% were identified as potential cellulose degraders. Within the phylum, there was a significant correlation between GH content and phylogeny (Mantel test, $P < 0.0001$) as clusters of phylogenetically related bacteria displayed homogeneous gene content (Fig. 4). Roughly, *Propionibacteriaceae* contained a low frequency of GH involved in the cellulose utilization, whereas organisms affiliated with *Streptomycineae*, although highly variable, contained on average 10 orthologous genes for possible β-glucosidase and four genes for hypothetical cellulase (Fig. 4 and 5). Thus, despite extensive variation in GH content, it appeared that the potential for cellulose utilization and the lack of potential were partially conserved within *Actinobacteria*.

The distribution of the three functional types at different taxonomic ranks was investigated following the complete taxonomic identification of each strain of the *Actinobacteria*. The three strategies were observed in the 514 strains of the class *Actinobacteria*. But when lower taxonomic ranks were considered, the proportion of taxa composed of bacteria having similar strategies increased (see Fig. S4 in the supplemental material). However, taxa with bacteria displaying different strategies were still observed at lower ranks, as 16% and 3% of the 96 species and 277 subspecies analyzed displayed two or more described functional types, respectively. We assumed that these low-level variations resulted from gene gain and loss events.

Lifestyles. We observed that the phylogeny and the GH genomic content of the bacteria were partially correlated, at least at lower taxonomic ranks. To identify possible mechanisms for this variation, we examined how bacterial lifestyle might affect the frequency of the GH types (β-glucosidases and cellulases). We first observed that facultative autotrophic organisms possessed fewer GH genes than heterotrophic organisms. Also, parasitic lineages such as *Legionella*, *Vibrio*, *Yersinia*, *Salmonella*, and *Mycobacterium* were analyzed, but these lineages generally contained few GH (Fig. 5). Thus, these variations in lifestyle may explain some of the phylogenetic variation in GH content that we observed in bacteria.

DISCUSSION

For decades, linking microbial community composition and function has been challenging. Nowadays, however, the availability of an increasing amount of sequenced genomes and consistent annotation systems provides one way to link phylogeny and the distribution of functional traits. Based on this idea, the phylogenetic distribution of 21,985 enzymes-genes related to the cellulose utilization in 5,123 sequenced bacterial genomes is identified here. This information has greatly expanded our knowledge of which microbial lineages have the metabolic potential to utilize cellulose and the byproducts of cellulose degradation. Further, we have observed extensive variation in this capability within each phylum. Our analysis of sequenced microbial genomes reveals that 20% of the analyzed strains have no relevant enzymes and are unlikely to utilize β-1,4 glucoside bond molecules as a carbon or energy source. A part of this group consists of parasites and facultative autotrophs. In contrast, we find that the majority (56%) of prokaryotes possess genes for enzymes allowing them to feed only on small oligosaccharides and cellobiose. These opportunists can potentially be considered "cheaters" (3), as they may take advantage of the degraders that produce exo- and endocellulases (24%).

We observe extensive interphylum diversity of the GH content, but the phylogenetic distribution of GH in sequenced genomes is clustered nonrandomly and mostly follows a Brownian motion-type evolution (i.e., species that are more closely related tend to have traits that are more similar). β-Glucosidases are widely distributed and form clades roughly at the genus level. In addition, putative cellulase-producing organisms form clusters of organisms having a 16S rRNA similarity equivalent to a common definition of a microbial "species." The clustering level of β-glucosidase-producing bacteria is supported by a previously independently obtained value for the *in vivo* ability of 738 bacterial strains to use cellobiose as a sole carbon source (23). As an example of this clustering, cellobiose utilization is common in *Actinobacteria*, but only a few clades within this phylum are putatively capable of cellulose degradation. This

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

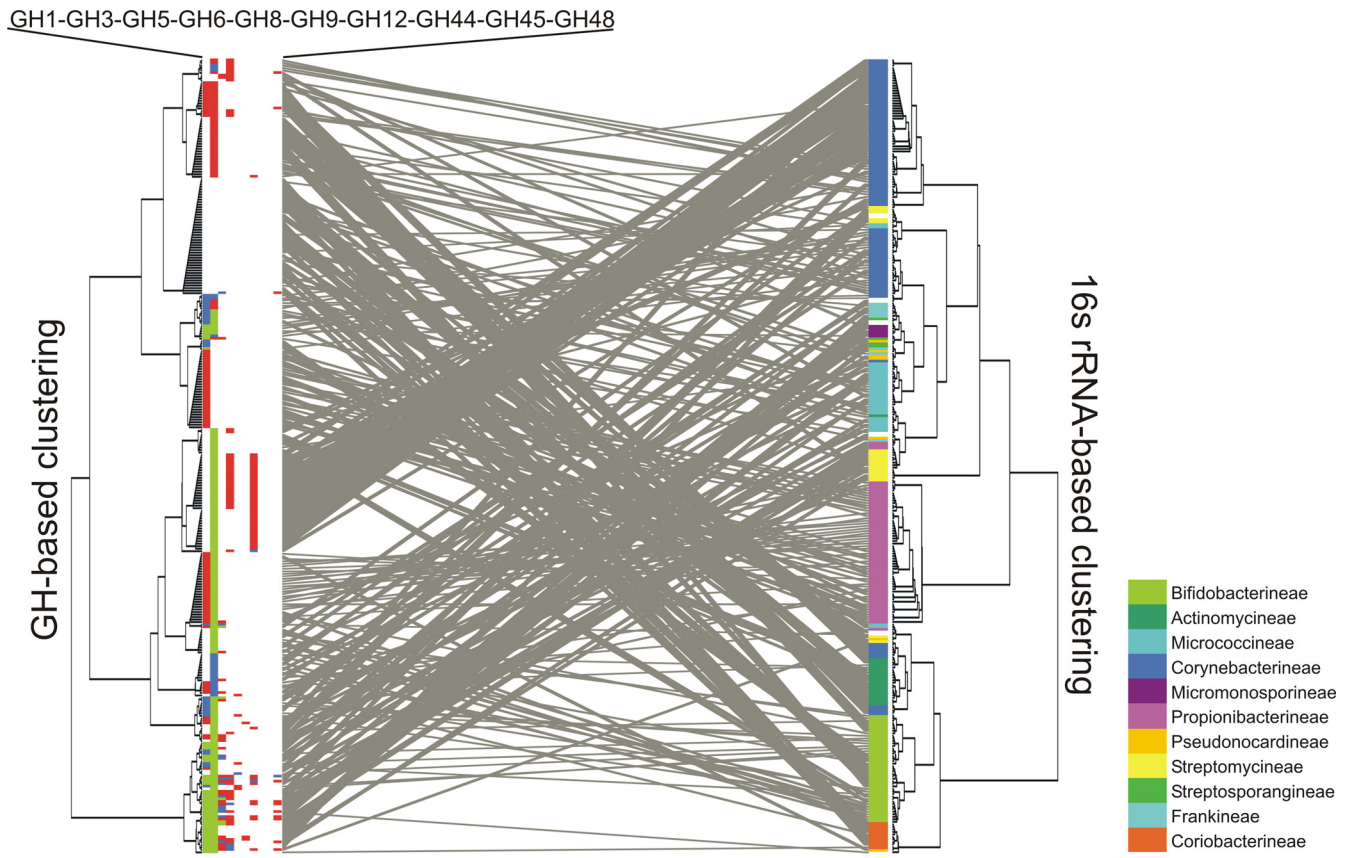


FIG 4 Comparison of the clustering of *Actinobacteria* based on GH content (left) and 16S rRNA phylogeny (right). The consensus phylogenetic tree was based on a 16S rRNA alignment retrieved from the SILVA database and constructed using Phylip (distance matrix F84, neighbor joining, 100 bootstraps). GH clustering was based on a Bray-Curtis dissimilarity index measured for the pairwise comparison of the glycoside hydrolase families involved in cellulose degradation of each sequenced genome (1 gene, blue; 2 genes, red; >2 genes, green). The gray lines connect identical strains in the displayed clustering.

clustered phylogenetic distribution of a putative cellulose degradation potential in bacteria is an important parameter for understanding the linkage between microbial community composition and cellulose degradation. A minority of lineages are described as potential cellulose degraders in our database. However, the abundance of degraders may have an important structuring effect on the overall community due to the large contribution of plant polymers to the overall carbon pool in soil ecosystems (2).

The analysis of the GH distribution also reveals what appears to be a redundancy of seemingly similar predicted functions for both β-glucosidase and cellulase in microorganisms. Putative cellulose degraders have between 1.2 and 2.7 more β-glucosidase genes from GH1 and GH3 than from their noncellulolytic relatives. The end product of the cellulose hydrolysis by endo- and exocellulases is cellobiose. This molecule is a glucose disaccharide having multiple properties: it is (i) a repressor of the cellulolytic machinery (44) and (ii) an inhibitor of the cellulases (45). A model for cellulose deconstruction suggests that the cellobiose concentration can increase rapidly during the cellulolysis and inhibit cellulase activity. Thus, increasing the β-glucosidase activity by having multiple enzymes may facilitate a more efficient cellulose degradation. This process has been demonstrated both *in vitro* and *in vivo* (45, 46). Alternatively, these seeming similar putative β-glucosidases may have slightly differentiated functions. These can include distinct substrate specificities, biochemical properties (e.g., pH_{opt}, T_{opt}), and regulation. Indeed, the plant cell wall deconstruction, as a whole, is considered a complex process requiring many enzymes. Their expression is assumed to be a response to the availability of specific substrates (44, 47).

We also observe redundancy in GH genes encoding putative polymer cellulose degradation. In our data set, potential cellulose degraders represent 24% of the analyzed strains, and 50% of these possess a plurality of genes for cellulases. Unfortunately, thorough and integrated characterizations of the cellulolytic systems expressed by fully sequenced plant cell wall degraders are unevenly represented. However, there is some evidence that an increase in GH diversity and abundance results in improved cellulose utilization abilities (1, 48). Thus, we speculate that a presence of seemingly redundant genes allows the bacteria to remain active in a wide range of environmental conditions (pH, substrate availability, temperature, etc.).

The idea of the widespread presence of genes encoding putative β-glucosidases is supported by the sequencing of several genomes/metagenomes (9, 49–52). In this report, we show that genes for GH1 and -3 are also abundant in most of the microbial genomes. Many environmental studies have measured β-glucosidase activity as a way to understand the potential of microbial communities to degrade plant polymers (53, 54). However, we show that traits having potential β-glucosidase function are very common among

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

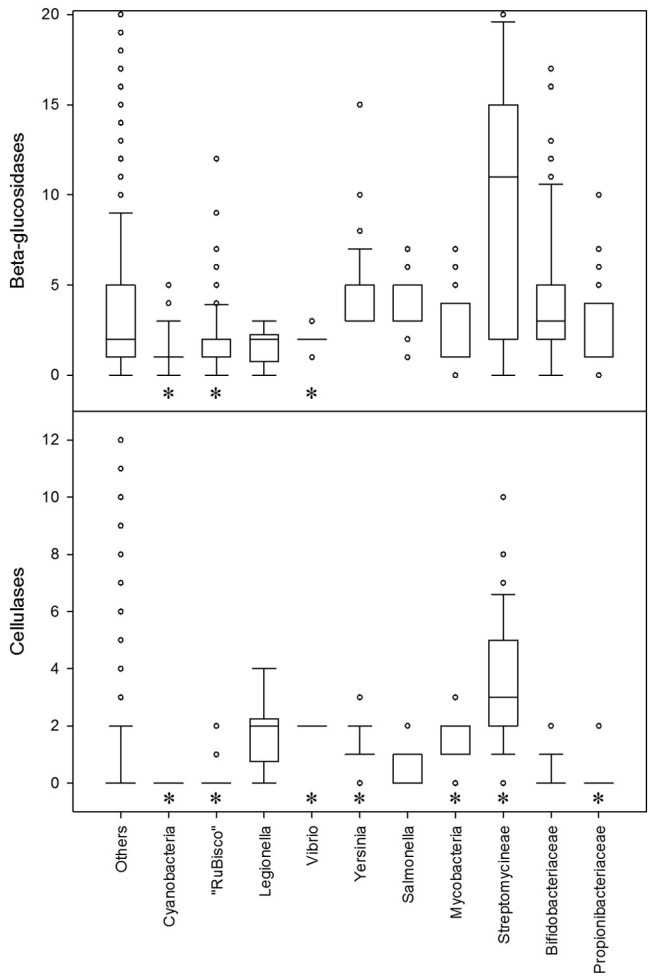


FIG 5 Abundance of putative β-glucosidases and cellulases in specific bacterial subpopulations, based on the presence of lifestyle-specific genetic markers, and in some *Actinobacteria* subgroups. The subpopulations were compared to the "Others" bacteria using the Welch *t* test (∗, $P < 1.10^{-16}$).

microbial lineages—especially in comparison to cellulose degradation potential—and many noncellulolytic organisms contain a putative β-glucosidase. This result is consistent with the observation that β-glucosidase activity in soil commonly is stable and high (54, 55). It suggests that β-glucosidase activity is common in most ecosystems and therefore a limited indicator of the cellulytic potential of a microbial community.

It is worth noting that our data set may be biased toward organisms that are involved in diseases. However, recent sequencing efforts have expanded the representation of non-disease-related microorganisms and our data set covers 28 phyla. In less-characterized phyla, the detection of GH may be difficult (56). Thus, our data likely represent a lower bound of the distribution of the considered GH families in nature. It is also important to recognize that not all glycoside hydrolase genes analyzed here encode proteins involved in the degradation of cellulose or oligosaccharides (15). Based on our analysis, at least 75% of the cellulases from GH8 are involved in the cellulose production in *Proteobacteria*. These cellulases may be involved in a transglycosylation reaction but have hydrolytic activity on cellulose analogs (e.g., carboxymethyl cellulose [CMC]) (31, 39, 57) and are thus frequently identified as hydrolytic cellulases. Although we have been careful in identifying putative biosynthetic enzymes, it is possible that other genes may have been misclassified, as some of these enzymes have been located outside the BCS operon (31) and/or as part of other GH families (39, 40).

In addition, less than 600 proteins from the GH families considered here have been biochemically confirmed. Thus, the functional annotation of most of the 21,985 presented traits remains hypothetical. However, it seems unlikely that new functional evidence would change our general findings on the phylogenetic distribution of the potential for cellulose utilization.

Interestingly, the glycoside hydrolases involved in the chitin utilization (i.e., chitinase and *N*-acetyl-glucosaminidase) display similar clustering patterns. Indeed, chitinases active on the polymeric substrate are found in small and highly clustered clades (>98% 16S rRNA similarity), whereas *N*-acetylglucosaminidases are identified in larger clusters (<98% 16S rRNA similarity) (24). Together with our findings, this suggests that genes involved in the processing of short and easy-to-metabolize substrates (e.g., cellobiose, *N*-acetylglucosamine) are broadly distributed in bacterial lineages, unlike the genes for polymer degradation. This also reveals that the vast majority of the bacterial lineage is unable to interact directly with the two most abundant polymers present on Earth and rely on enzyme producers to release short substrates. Nevertheless, linking the natural microbial community structure and the process of polymer deconstruction is a complex task. However, it is assumed that the fitness of each environmental strain is strongly associated with its strategy regarding polymer utilization (opportunist, degrader), with the distribution of the strategies of others, and with the qualitative variability of resources (6).

Little is known about the vast majority of the naturally occurring bacteria regarding their ability to interact with plant-derived material (58). Thus, as presented here, connecting the functional traits to phylogenetically described bacterial lineages in a phylogenomic framework is a valuable input for ecosystem modeling. Our approach discriminates the phylogenetic levels of the functional status: the potential cellulose degraders and the potential opportunists are not clustered at the same 16S rRNA distance. Further, we provide information on the functional status of each bacterial lineage/taxon regarding its potential for cellulose utilization. Thus, given the knowledge of the composition of microbial consortia (e.g., 16S rRNA diversity and abundance), our general framework can facilitate an identification of the cellulolytic potential of environmental communities. This is important for understanding the functioning of the microbial soil ecosystem, as cellulose deconstruction is a key aspect of the C cycling.

ACKNOWLEDGMENTS

We thank Steven Allison, Jennifer Martiny, Amy Zimmerman, and Kathleen Treseder for many helpful comments on the manuscript.

This work was supported by the DE-PS02-09ER09-25 grant from the U.S. Department of Energy.

REFERENCES

1. **Wilson DB.** 2011. Microbial diversity of cellulose hydrolysis. Curr. Opin. Microbiol. **14**:259–263.
2. **Goldfarb KC, Karaoz U, Hanson CA, Santee CA, Bradford MA, Treseder KK, Wallenstein MD, Brodie EL.** 2011. Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. Front. Microbiol. **2**:94. doi:10.3389/fmicb.2011.00094.

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

3. **Allison SD.** 2005. Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. Ecol. Lett. **8**:626–635.
4. **Romaní AM, Fischer H, Mille-Lindblom C, Tranvik LJ.** 2006. Interactions of bacteria and fungi on decomposing litter: differential extracellular enzyme activities. Ecology **87**:2559–2569.
5. **Allison SD, Martiny JB.** 2008. Colloquium paper: resistance, resilience, and redundancy in microbial communities. Proc. Natl. Acad. Sci. U. S. A. **105**(Suppl 1):11512–11519.
6. **Allison SD.** 2012. A trait-based approach for modelling microbial litter decomposition. Ecol. Lett. **15**:1058–1070.
7. **Moorhead DL, Sinsabaugh RL.** 2006. A theoretical model of litter decay and microbial interaction. Ecol. Monogr. **76**:151–174.
8. **Gilbert HJ.** 2010. The biochemistry and structural biology of plant cell wall deconstruction. Plant Physiol. **153**:444–455.
9. **Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B.** 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. **37**:D233–D238.
10. **Blake AW, McCartney L, Flint JE, Bolam DN, Boraston AB, Gilbert HJ, Knox JP.** 2006. Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes. J. Biol. Chem. **281**:29321–29329.
11. **Cuskin F, Flint JE, Gloster TM, Morland C, Basle A, Henrissat B, Coutinho PM, Strazzulli A, Solovyova AS, Davies GJ, Gilbert HJ.** 2012. How nature can exploit nonspecific catalytic and carbohydrate binding modules to create enzymatic specificity. Proc. Natl. Acad. Sci. U. S. A. **109**:20889–20894.
12. **Deutscher J, Francke C, Postma PW.** 2006. How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. Microbiol. Mol. Biol. Rev. **70**:939–1031.
13. **Schlösser A, Aldekamp T, Schrempf H.** 2000. Binding characteristics of CebR, the regulator of the ceb operon required for cellobiose/cellotriose uptake in Streptomyces reticuli. FEMS Microbiol. Lett. **190**:127–132.
14. **Gabor E, Gohler AK, Kosfeld A, Staab A, Kremling A, Jahreis K.** 2011. The phosphoenolpyruvate-dependent glucose-phosphotransferase system from Escherichia coli K-12 as the center of a network regulating carbohydrate flux in the cell. Eur. J. Cell Biol. **90**:711–720.
15. **Mba Medie F, Davies GJ, Drancourt M, Henrissat B.** 2012. Genome analyses highlight the different biological roles of cellulases. Nat. Rev. Microbiol. **10**:227–234.
16. **Davies G, Henrissat B.** 1995. Structures and mechanisms of glycosyl hydrolases. Structure **3**:853–859.
17. **Vuong TV, Wilson DB.** 2010. Glycoside hydrolases: catalytic base/nucleophile diversity. Biotechnol. Bioeng. **107**:195–205.
18. **Davies GJ, Gloster TM, Henrissat B.** 2005. Recent structural insights into the expanding world of carbohydrate-active enzymes. Curr. Opin. Struct. Biol. **15**:637–645.
19. **Aspeborg H, Coutinho PM, Wang Y, Brumer H, III, Henrissat B.** 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evol. Biol. **12**:186. doi:10.1186/1471-2148-12-186.
20. **Henrissat B.** 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **280**(Pt 2):309–316.
21. **Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC.** 2010. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology **20**:1574–1584.
22. **Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y.** 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. **40**:W445–W451.
23. **Martiny AC, Treseder KK, Pusch GD.** 13 December 2012. Phylogenetic conservatism of functional traits in microorganisms. ISME J. [Epub ahead of print.] doi:10.1038/ismej.2012.160.
24. **Zimmerman AE, Martiny AC, Allison SD.** Microdiversity of extracellular enzyme genes among sequenced prokaryotic genomes. ISME J., in press.
25. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD.** 2012. The Pfam protein families database. Nucleic Acids Res. **40**:D290–D301.
26. **Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V.** 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. **33**:5691–5702.
27. **Disz T, Akhter S, Cuevas D, Olson R, Overbeek R, Vonstein V, Stevens R, Edwards RA.** 2010. Accessing the SEED genome databases via Web services API: tools for programmers. BMC Bioinformatics **11**:319. doi:10.1186/1471-2105-11-319.
28. **Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Stevens RL, Vonstein V, Xia F.** 2012. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. PLoS One **7**:e48053. doi:10.1371/journal.pone.0048053.
29. **Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW.** 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect. Immun. **79**:4286–4298.
30. **Kumar Y, Valdivia RH.** 2009. Leading a sheltered life: intracellular pathogens and maintenance of vacuolar compartments. Cell Host Microbe **5**:593–601.
31. **Römling U.** 2002. Molecular biology of cellulose production in bacteria. Res. Microbiol. **153**:205–212.
32. **Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. **35**:7188–7196.
33. **Retief JD.** 2000. Phylogenetic analysis using PHYLIP. Methods Mol. Biol. **132**:243–258.
34. **Felsenstein J.** 2006. PHYLIP (Phylogeny Inference Package), 3.65 ed. Department of Genome Sciences, University of Washington, Seattle, WA.
35. **Letunic I, Bork P.** 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. **39**:W475–W478.
36. **Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHN, Wagner H.** 2012. Package 'vegan'—Community Ecology Package, version 2.0-4. http://cran.r-project.org/web/packages/vegan/vegan.pdf.
37. **Paradis E, Claude J, Strimmer K.** 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20**:289–290.
38. **Fritz SA, Purvis A.** 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conserv. Biol. **24**:1042–1051.
39. **Berlemont R, Delsaute M, Pipers D, D'Amico S, Feller G, Galleni M, Power P.** 2009. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. ISME J. **3**:1070–1081.
40. **Xu H, Chater KF, Deng Z, Tao M.** 2008. A cellulose synthase-like protein involved in hyphal tip growth and morphological differentiation in Streptomyces. J. Bacteriol. **190**:4971–4978.
41. **Mba Medie F, Ben Salah I, Drancourt M, Henrissat B.** 2010. Paradoxical conservation of a set of three cellulose-targeting genes in Mycobacterium tuberculosis complex organisms. Microbiology **156**:1468–1475.
42. **Yan Y, Yang J, Dou Y, Chen M, Ping S, Peng J, Lu W, Zhang W, Yao Z, Li H, Liu W, He S, Geng L, Zhang X, Yang F, Yu H, Zhan Y, Li D, Lin Z, Wang Y, Elmerich C, Lin M, Jin Q.** 2008. Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated Pseudomonas stutzeri A1501. Proc. Natl. Acad. Sci. U. S. A. **105**:7564–7569.
43. **Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D.** 2007. Genomics of actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiol. Mol. Biol. Rev. **71**:495–548.
44. **Zhang YH, Lynd LR.** 2005. Regulation of cellulase synthesis in batch and continuous cultures of Clostridium thermocellum. J. Bacteriol. **187**:99–106.
45. **Bommarius AS, Katona A, Cheben SE, Patel AS, Ragauskas AJ, Knudson K, Pu Y.** 2008. Cellulase kinetics as a function of cellulose pretreatment. Metab. Eng. **10**:370–381.
46. **Gefen G, Anbar M, Morag E, Lamed R, Bayer EA.** 2012. Enhanced cellulose degradation by targeted integration of a cohesin-fused beta-

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest

glucosidase into the Clostridium thermocellum cellulosome. Proc. Natl. Acad. Sci. U. S. A. **109**:10298–10303.
47. **Langlois P, Bourassa S, Poirier GG, Beaulieu C.** 2003. Identification of Streptomyces coelicolor proteins that are differentially expressed in the presence of plant material. Appl. Environ. Microbiol. **69**:1884–1889.
48. **Fontes CM, Gilbert HJ.** 2010. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. Annu. Rev. Biochem. **79**:655–681.
49. **Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM.** 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science **331**:463–467.
50. **Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernandez M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR.** 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature **450**:560–565.
51. **Zhu L, Wu Q, Dai J, Zhang S, Wei F.** 2011. Evidence of cellulose metabolism by the giant panda gut microbiome. Proc. Natl. Acad. Sci. U. S. A. **108**:17714–17719.
52. **Delmont TO, Malandain C, Prestat E, Larose C, Monier JM, Simonet P, Vogel TM.** 2011. Metagenomic mining for microbiologists. ISME J. **5**:1837–1843.
53. **Lammirato C, Miltner A, Wick LY, Kastner M.** 2010. Hydrolysis of cellobiose by beta-glucosidase in the presence of soil minerals—interactions at solid-liquid interfaces and effects on enzyme activity levels. Soil Biol. Biochem. **42**:2203–2210.
54. **Sinsabaugh RL, Lauber CL, Weintraub MN, Ahmed B, Allison SD, Crenshaw C, Contosta AR, Cusack D, Frey S, Gallo ME, Gartner TB, Hobbie SE, Holland K, Keeler BL, Powers JS, Stursova M, Takacs-Vesbach C, Waldrop MP, Wallenstein MD, Zak DR, Zeglin LH.** 2008. Stoichiometry of soil enzyme activity at global scale. Ecol. Lett. **11**:1252–1264.
55. **Nannipieri P, Kandeler E, Ruggiero P.** 2002. Enzyme activities and microbiological and biochemical processes in soil, p 1–33. *In* Burns RG, Dick RP (ed.), Enzymes in the environment: activity, ecology and applications. CRC Press, Boca Raton, FL.
56. **Meyer F, Overbeek R, Rodriguez A.** 2009. FIGfams: yet another set of protein families. Nucleic Acids Res. **37**:6643–6654.
57. **Mazur O, Zimmer J.** 2011. Apo- and cellopentaose-bound structures of the bacterial cellulose synthase subunit BcsZ. J. Biol. Chem. **286**:17601–17606.
58. **Sukharnikov LO, Cantwell BJ, Podar M, Zhulin IB.** 2011. Cellulases: ambiguous nonhomologous enzymes in a genomic perspective. Trends Biotechnol. **29**:473–479.

Downloaded from http://aem.asm.org/ on August 29, 2018 by guest