Genomic Potential for Polysaccharide Deconstruction in Bacteria

Renaud Berlemont, Adam C. Martiny
Department of Biological Sciences, California State University Long Beach, Long Beach, California, USA; Department of Earth System Science and Department of Ecology and Evolutionary Biology, University of California at Irvine, Irvine, California, USA

Glycoside hydrolases are important enzymes that support bacterial growth by enabling the degradation of polysaccharides (e.g., starch, cellulose, xylan, and chitin) in the environment. Presently, little is known about the overall phylogenetic distribution of the genomic potential to degrade these polysaccharides in bacteria. However, knowing the phylogenetic breadth of these traits may help us predict the overall polysaccharide processing in environmental microbial communities. In order to address this, we identified and analyzed the distribution of 392,166 enzyme genes derived from 53 glycoside hydrolase families in 8,133 sequenced bacterial genomes. Enzymes for oligosaccharides and starch/glycogen were observed in most taxonomic groups, whereas glycoside hydrolases for structural polymers (i.e., cellulose, xylan, and chitin) were observed in clusters of relatives at taxonomic levels ranging from species to genus as determined by consenTRAIT. The potential for starch and glycogen processing, as well as oligosaccharide processing, was observed in 85% of the strains, whereas 65% possessed enzymes to degrade some structural polysaccharides (i.e., cellulose, xylan, or chitin). Potential degraders targeting one, two, and three structural polysaccharides accounted for 22.6, 32.9, and 9.3% of genomes analyzed, respectively. Finally, potential degraders targeting multiple structural polysaccharides displayed increased potential for oligosaccharide deconstruction. This study provides a framework for linking the potential for polymer deconstruction with phylogeny in complex microbial assemblages.

Together, carbohydrates account for ~75% of the earth’s biomass (1), and therefore, our understanding of how this material is processed by microorganisms is a key question for global biogeochemistry. Of these carbohydrates, the more abundant are cellulose, xylan, chitin, starch, and glycogen. Cellulose, the most abundant polysaccharide on earth, is produced by a variety of organisms, including plants and bacteria. Xylan and several other polysaccharides (e.g., glucomannan) together form hemicellulose, associated with cellulose, to constitute plant cell wall material. Chitin, produced by many arthropods (e.g., crustaceans and insects) and fungi, is the second most abundant structural polysaccharide. On the other hand, starch and glycogen, found in many organisms, are used as a means to store energy inside the cell. Additional important polysaccharides are dextrans and fructans (e.g., inulin). These polysaccharides represent the major source of energy for most heterotrophs, including many microbes (e.g., bacteria and fungi). These heterotrophs produce glycoside hydrolases (GHs) to degrade polysaccharides and to release monosaccharides (e.g., glucose, fructose, N-acetylglucosamine, and xylose) that are key resources for microbial growth (2–4). However, not all microorganisms are directly involved in polysaccharide deconstruction. Active degraders express enzymes (e.g., endo-/exocellulases and endo-/beta-xylanase) that act synergistically in the deconstruction of complex polymers (e.g., cellulose and xylan) and enzymes for processing small degradation products (5, 6). Many degraders are likely to be involved in the degrada
tion of multiple substrates since, in nature, polysaccharides form complex mixed materials (e.g., plant material) (7).

Some other lineages (i.e., opportunists) are solely capable of processing the smaller substrates (2). In addition, many members of the Tenericutes, Spirochaetes, or Cyanobacteria are unable to process cellulose or its deconstruction product (8).

Understanding the distribution of genes for polysaccharide utilization in bacteria is key to understanding how microbial communities, including bacteria and fungi, are assembled and evolve and, more importantly, how many ecosystems function (2, 9–11). However, to date, little is known about the distribution and the cooccurrence of the bacterial potentials for deconstruction of the variety of polysaccharides encountered in nature.

The GH superfamily encompasses proteins involved in glucan hydrolysis, including cellulose, hemicellulose, xylan, chitin, starch, and glycogen (12). The classification of glycoside hydrolases is primarily based on protein structures that common allow for Pfam (i.e., probabilistic models for statistical inference of homology) identification (13). Generally, some enzymes are active on polymers and generate oligosaccharides, which are subsequently processed by a variety of enzymes with activity on oligosaccharides (i.e., the “-osidas”) that are found in a few GH families (e.g., GH1 to -3). Considering their central function in carbon metabolism (14), their impact on nutrient cycling (15), and their biotechnological potential (16), the biochemistry of these enzymes has been widely investigated. In some cases, glycoside hydrolase families with different folds display similar activities (e.g., endocellulases from GH5 and -9), while, in some families, minor modifications of the active site/substrate-binding cleft result in different specificities/activities while conserving the overall fold (e.g., exocellulase and endocellulase from GH6). However, despite examples where minor modifications of the active site lead
to distinct substrate specificities (e.g., cellulases and chitosanases from GH8), the substrate specificities of many GH families are remarkably conserved (see Table S1 in the supplemental material). This suggests a strong selective pressure on key enzymes for polysaccharide deconstruction (8, 17–19).

In microorganisms, the phylogenetic distribution of enzymes is not random (8, 20), and the presence of glycoside hydrolases is restricted at different taxonomic levels. For example, as described in CAZy (12), cellulases from glycoside hydrolase family 7 are exclusively observed in members of the Fungi, suggesting an ancestral specialization event. In contrast, the distribution of various traits targeting cellulose and chitin has been shown to be broad and the corresponding GHs to be phylogenetically conserved in clusters at the genus or species level (8, 17). Functional status (e.g., potential degrader or opportunistic) has been proposed as a concept to discriminate lineages according to their potential for cellulose decomposition (8). However, cellulose in plant material is commonly associated with other polymers. Thus, an important question is whether or not potential cellulose degraders are also involved in hydrolysis of other plant polymers or are reliant on the properties of other organisms. This information, in association with taxonomic identification of microbial assemblages, is key in order to predict and understand how changes in microbial community composition in response to environmental perturbation (e.g., global climate change) would affect polysaccharide processing.

To improve our understanding of polysaccharide processing by bacteria, we asked the following questions: (i) what is the phylogenetic distribution of enzyme potentials for polysaccharide deconstruction in bacteria, and (ii) how are the enzyme potentials for deconstruction of different polysaccharides linked? To answer these questions, we analyzed the distribution of orthologs of 53 glycoside hydrolase families targeting various plant, fungal, bacterial, and animal polysaccharides in 8,133 sequenced bacterial genomes. Our results provide a basic framework for determining the potential for polysaccharide processing in taxonomically resolved bacterial assemblages.

Materials and Methods

Glycoside hydrolase mapping. To prevent the loss of potential GHs, we developed a custom bioinformatics pipeline based on a well-defined hidden Markov model (HMM) to detect potential GHs with low similarity to known enzymes. Linking this annotation with the genomic context, as provided by the FIGfam annotation, allowed us to further discriminate the functions of proteins belonging to the same GH family, as follows. (i) We extracted protein sequences for all SEED-annotated genes in fully sequenced bacterial genomes (see Table S2 in the supplemental material). The most frequent GHs with high potential for structural polysaccharide processing, including GH1, GH8, GH38, targeting cellulose, dextrans, fructans, rhamnose, and mannose, respectively.

Phylogenetic distribution of glycoside hydrolases. The GH content was extremely variable across phyla. For example, both Actinobacteria and Bacteroidetes displayed abundant and diverse GHs with high potential for structural polysaccharide processing, while, perhaps not surprisingly given that most are phototrophic, Cyanobacteria harbored reduced potentials for polysaccharide processing. In phyla with few sequenced genomes, such as Verrucomicrobia and Acidobacteria (16 and 5 analyzed genomes, respectively), the GH content was generally abundant but highly variable. At a finer taxonomic level, the distribution of GH families across lineages from the same phyla was also highly variable, as suggested by the comparison of GH distribution between the
mostly pathogenic *Mycobacterium* strains and the mostly environmental *Streptomyces* strains (phylum *Actinobacteria*) (see Fig. S1 in the supplemental material). When present, genes for GH families targeting starch and oligosaccharides displayed high intragenomic redundancy ($\sigma > 0.80$) compared to that of genes for GHs targeting structural polymers. Among these, some GH families displayed extremely low redundancy (e.g., GH8 [$\sigma < 0.10$]) (see Table S1). Thus, many strains had important potential for oligosaccharides and starch or glycogen processing. Conversely, GHs for structural polysaccharide deconstruction were less frequent or, when present, less abundant.

Next, we investigated the phylogenetic distribution of each individual GH family in the bacterial genomes using metrics for phylogenetic signal (D) and trait depth ($\tau_D$). For most of the GH families, the phylogenetic signal D, either positive or negative, was close to zero, revealing clumped (Brownian) distributions (Fig. 2A). Conversely, GH families 71, 75, 44, and 45 displayed D values corresponding to more dispersed traits. On average, most
of the GH families were observed in bacteria forming clusters of relatives with a clade depth $D$ of less than 0.03 for 16S rRNA gene distance (i.e., sharing of traits was more common in organisms sharing 94% or more 16S rRNA sequence similarity). Few GHs, mostly the abundant ones, like GH1 and 13, displayed higher $D$ values. $D$ and $D'$ were also computed for functional groups. As described for individual GH families, the phylogenetic signal $D$ for functional groups targeting starch and glycogen, cellulose, fructose, and dextran was positive and close to 0 ($D \approx 0.1$). In contrast, the potential to target oligosaccharides and xylan displayed a more dispersed distribution ($D = 0.24$ and $D = 0.17$, respectively). The corresponding $D'$ values were also estimated. On average, the potential to target xylan, dextran, or fructan was observed in bacteria forming clusters of relatives with a clade depth $D'$ of below 0.025 for 16S rRNA gene distance. The potential to target other substrates was observed in groups of relatives forming clusters with $D'$ values of $>0.80$. For these functional potentials, we computed the $D'$ of strains lacking these functional groups. On average, strains having no detected potential for processing of oligosaccharides, starch, cellulose, and chitin formed clusters of bacteria with average $D'$ values equal to 0.01, 0.01, 0.02, and 0.02, respectively.

Since strains that are closely related according to their 16S rRNA sequences are assumed to share increased numbers of functional genes, we then investigated how phylogeny and the overall genome-specific glycoside hydrolase content were linked (Fig. 3; see also Table S3 in the supplemental material). Strains forming operational taxonomic units with $100, \geq 99, \geq 97.5,$ and $\geq 95\%$ 16S rRNA gene similarities displayed median functional dissimilarities of 0.062, 0.100, 0.183, and 0.308, respectively. This suggested that even very closely related strains may differ regarding their GH composition and, thus, may have distinct potential regarding polysaccharide deconstruction. Strains with lower 16S rRNA gene sequence similarity displayed further increases in functional dissimilarity (median overall functional dissimilarity of 0.733).

**Cooccurrence of potential for polysaccharide processing.** Next, we examined the cooccurrence of potentials for polysaccharide hydrolysis across organisms (Fig. 4). First, 3.1% of the strains analyzed had genes for oligosaccharide processing only (i.e., α-oligosaccharides), whereas 88.3% displayed the combined potential to process oligosaccharides and starch/glycogen. The ability to target at least one structural polysaccharide (i.e., cellulose, xylan, or chitin) was observed in 66.7% of the genomes (Fig. 4A), and the potential for cellulose, xylan, and chitin deconstruction was detected in 40.8, 24.6, and 53.1% of sequenced bacterial genomes, respectively. Most of the strains classified as potential degraders contained genes for the hydrolysis of multiple targets, with 32.9 and 9.4% having the potential to target 2 or 3 structural polysaccharides, respectively, and only 24.4% of the genomes having the potential to target a single structural polysaccharide. Most of the potential chitin degraders (74.8%) also displayed the potential to...
target plant structural polysaccharides. Of the potential cellulose degraders, 29.5 and 76.9% also targeted xylan and chitin, while 48.9 and 71.8% of the potential xylan degraders could also target cellulose and chitin, respectively. A small percentage of the strains, 5.1%, had no detected enzymes for oligosaccharide hydrolysis but had enzymes for starch/glycogen, structural polysaccharides, or other substrates (i.e., dextran, fructan, or other plant or animal polymers) or multiple substrates (i.e., mixed). Cellulases from GH8 for cellulose biosynthesis (BcsZ) were observed in 1,686 genomes (mostly Proteobacteria); among these, 476 genomes also possessed the potential for cellulose deconstruction.

Next, we investigated the link between the potential for structural polysaccharide and oligosaccharide processing. Strains with the potential to target multiple structural polysaccharides displayed significantly more enzymes for oligosaccharides than strains with reduced potential (Fig. 4B). Finally, we investigated the correlation between potentials for polysaccharide deconstruction (see Table S4 in the supplemental material). The abundance of genes for enzymes targeting cellulose was positively correlated with the abundance of genes for most of the other functions, including the potential for xylan and chitin deconstruction. So were most of the potentials for polysaccharides hydrolysis. However, cellulose biosynthesis and dextran deconstruction were negatively correlated with the potential for cellulose deconstruction.

**DISCUSSION**

Linking phylogeny to function is a recurrent question in microbiology and is key to understanding how changes in microbial communities in response to environmental perturbation would affect ecosystem functioning (e.g., carbon cycling). The increasing number of sequenced bacterial genomes annotated using consistent techniques is enabling a more thorough assessment of the patterns of distribution of functions across the microbial tree of life. Recently, a systematic investigation of sequenced bacterial genomes for functional traits related to cellulose (8) or chitin (17) deconstruction has improved our understanding of the evolutionary forces shaping the present distribution of potentials for cellulose and chitin deconstruction in sequenced bacterial genomes (33). Understanding the distribution of cellulases in bacteria allows the discrimination of potential cellulose degraders, opportunists, and bacteria not involved in cellulose deconstruction. This information can be used as a basic framework to identify and investigate the distribution of functional groups, e.g., cellulose degraders, in environmental microbial assemblages (34–36). In nature, cellulose is associated with various polymers (e.g., xylan) in most plant materials (5). Starch and glycogen are found in many organisms and are also important carbon sources for many bacteria (37), whereas chitin produced by fungi and arthropods is present in many environments (38). Finally, dextran and fructan (i.e., inulin) are also frequently observed polymers. Thus, understanding how bacteria access these resources is a key prerequisite to understanding the functioning of many ecosystems.

Here, we present an integrated phylogenomic analysis of the distribution of 392,166 sequences from the major GH families targeting the most abundant polysaccharides in 8,133 sequenced bacterial genomes. On average, there are 48.2 GHs/genome. Considering an average bacterial genome of ~3 Mbp and an average gene size of ~1.5 kbp (39), this represents 2.4% of the genes in bacteria. This value is in good agreement with previously assumed glycoside hydrolase frequency in bacterial genomes (40). However, GHs display broad distribution and most glycoside hydrolase families, being nonrandomly distributed, are found in smaller clusters of relatives at the genus or species level. The functional dissimilarity between strains increases with the phylogenetic distance, although closely related strains can display substantial functional divergences. This suggests that some recent events may have affected some genomes (e.g., gene duplication and horizontal gene transfer). Nevertheless, most of the traits analyzed are predominantly inherited from parent strains.

Knowing the phylogenetic distribution of the GHs, with some information regarding their substrate specificity, allowed us to expand the terminology introduced for cellulose degradation (i.e.,
opportu

display increased potential for polysaccharide deconstruction in order to cope with changing environments and fluctuating carbon sources. However, pathogens (e.g., Mycobacterium strains), relying on their hosts to provide resources, tend to have reduced potential for polysaccharide deconstruction.

Globally, we provide herein an integrated framework for potential polysaccharide deconstruction that connects glycoside hydrolases to functional status (e.g., potential xylan degrader) and to the phylogeny of sequenced bacterial lineages. This information is key in order to estimate the potential for polysaccharide deconstruction in microbial populations dominated by bacteria (49) and to parameterize simulated microbial guilds for ecosystem modeling (50). More broadly, this information will be useful to investigate how fluctuations of environmental bacterial communities, together with fungal populations (9), in response to natural or anthropogenic changes may affect the polysaccharide deconstruction in the environment and, more globally, the carbon cycling.

ACKNOWLEDGMENTS

We thank Eoin L. Brodie for many helpful comments on the manuscript. This material is based upon work supported by the NSF Dimensions of Biodiversity program (OCE-1046297) and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER), under award number DE-PS02-09ER00025.

The authors declare no conflict of interest.

REFERENCES


