SCIENTIFIC REPORTS

OPEN

GeneHunt for rapid domain-specific annotation of glycoside hydrolases

S. N. Nguyen[1], A. Flores[1], D. Talamantes[1], F. Dar[1], A. Valdez[1], J. Schwans[2] & R. Berlemont[1]

Received: 13 December 2018
Accepted: 26 June 2019
Published online: 12 July 2019

The identification of glycoside hydrolases (GHs) for efficient polysaccharide deconstruction is essential for the development of biofuels. Here, we investigate the potential of sequential HMM-profile identification for the rapid and precise identification of the multi-domain architecture of GHs from various datasets. First, as a validation, we successfully reannotated >98% of the biochemically characterized enzymes listed on the CAZy database. Next, we analyzed the 43 million non-redundant sequences from the M5nr data and identified 322,068 unique GHs. Finally, we searched 129 assembled metagenomes retrieved from MG-RAST for environmental GHs and identified 160,790 additional enzymes. Although most identified sequences corresponded to single domain enzymes, many contained several domains, including known accessory domains and some domains never identified in association with GH. Several sequences displayed multiple catalytic domains and few of these potential multi-activity proteins combined potentially synergistic domains. Finally, we produced and confirmed the biochemical activities of a GH5-GH10 cellulase-xylanase and a GH11-CE4 xylanase-esterase. Globally, this "gene to enzyme pipeline" provides a rationale for mining large datasets in order to identify new catalysts combining unique properties for the efficient deconstruction of polysaccharides.

Glycoside Hydrolases (GHs) are Carbohydrate-Active Enzymes (CAZy) that catalyze the hydrolysis of the glycosidic linkage in polysaccharides (e.g., cellulose, chitin) and oligosaccharides (e.g., cellobiose, chitobiose)[1]. GHs are found as single domain proteins (SDGHs) or associated with accessory domains such as carbohydrate binding modules (CBMs) within multi-domain GHs (MDGHs)[2]. In MDGHs, CBMs enhance enzyme-substrate interaction by anchoring the catalytic domain to the substrate[3]. The anchoring reduces diffusion from the substrate and locally increases the concentration of catalytic domains[4], thus improving the overall polysaccharide degradation[3].

GHs support essential processes for ecosystem function and for biotechnology. Among others, in land ecosystems, the deconstruction of plant biomass by microbial GHs is essential[5,6], whereas the breakdown of chitin, from arthropods and fungi, is important in both marine[7,8] and terrestrial ecosystems[9–12]. Next, in the gut of animals, microbial GHs target polysaccharides, supplement the lack of endogenous enzymes[13,14] and thus contribute to the processing of complex carbohydrates during digestion[15–18]. Finally, GHs are essential for the biofuel industry, as plant based polysaccharides constitute a major source of sustainable and renewable material capable of providing liquid transportation fuel[19–22].

Many GH-genes and proteins have been identified in a growing number of sequenced genomes and environmental samples thanks to the use of activity-driven screening[23,24] and bioinformatic annotation systems[12,16,18,25]. The precise identification of GH-genes and proteins is essential in order to understand how microbes support key functions across ecosystems[17,25,26] and to identify new enzymes for biotechnological application[18,21,27].

In order to identify new catalysts for biomass degradation, we examined the performance of sequential Hidden Markov Model (HMM) identifications[28] combined with publicly accessible HMM-profiles from the PFam database[29], here referred to as the GeneHunt approach[2,30], to detect GH-sequences and investigate their detailed architecture (i.e., the precise domain organization of MDGHs)[2]. More precisely, we first validated the GeneHunt approach by re-annotating the biochemically characterized GHs listed on the CAZy database (as of June 2018)[1]. As described for cellulases, xylanases, and chitinases[30], we expected the PFam-based annotation to correctly identify most of the proteins from the major GH families, although rare and recently identified GH-families would display inconsistencies. Next, we identified GHs in the M5nr database (version 13.12.15) containing 43,098,145 non-redundant, mostly microbial, protein sequences[31]. This collection of sequences derived from the major sequence database serves as the reference database for the MG-RAST annotation pipeline[31,32]. Finally, we identified the detailed multi-domain architecture of GH proteins in assembled, publicly accessible, metagenomes from

[1]Department of Biological Sciences, California State University Long Beach, Long Beach, California, USA. [2]Department of Chemistry and Biochemistry, California State University Long Beach, Long Beach, California, USA. Correspondence and requests for materials should be addressed to R.B. (email: Renaud.berlemont@csulb.edu)

MG-RAST. We hypothesized that, across database, GH proteins would exist primarily as single domain enzymes with low frequency of MDGHs as identified in sequenced microbial genomes[2,10]. We also expected that identified MDGHs would mostly consist in association between GH domains and CBMs as identified in the CAZy database[1]. Among the MDGHs, we expected to identify proteins with multiple catalytic domains and identified as potential multi-activity GHs (MAGHs). These MAGHs would display multiple catalytic domains with potential synergistic activities. In these proteins, synergistic interactions between catalytic domains would result from the complementarity of the associated domains and from a proximity effect[3]. More precisely, we envisioned (i) parallel pathway synergy where the combined catalytic domains target distinct, yet physically associated, substrates (e.g., cellulase:xylanase) and (ii) debranching synergy where one catalytic domain cleaves the side groups in substituted polysaccharides or cleaves branch points in reticulated polysaccharides, thus increasing the accessibility of the polysaccharide backbone for the second catalytic domain (e.g., xylanase:xylan-esterase). Relative to SDGH and MDGH (with CBMs), these MAGHs represent vastly untapped enzymatic diversity with great potential for improved biomass deconstruction[21,33–35].

Globally, this work provides the rationale and validation for the detailed and rapid detection of most identified GH families and associated domains from a variety of datasets, ranging from biochemically-characterized enzymes, the largest non-redundant sequence database, and assembled metagenomes derived from various environments. In addition, this work provides an exhaustive list of domains associated with GH domains, explores the diversity of multi-domain and multi-activity GHs, and identifies new types of catalysts with potential for biotechnological application.

Results

Mapping of glycoside hydrolases in CAZy and M5nr. First, in order to evaluate the GeneHunt approach, we (re)annotated the sequences of biochemically characterized GHs listed on the CAZy database. GeneHunt consistently annotated 7,620 GH sequences, out of 7,920 tested proteins (Table S1, Supplementary Data 1). For example, among the 327 biochemically characterized GH1s retrieved from the CAZy database, 325 (99.39%) of the sequences matched with PF00232 (i.e., "Glyco_hydro_1") whereas the 2 mis-annotated GH1s were short fragments of sequences identified in cDNA libraries with biochemical characterization remaining elusive to date (e.g., myrosinase from *Sinapis alba*, CAA42536.1). Likewise, most GH families listed on the CAZy database were consistently identified using the GeneHunt approach (Table S1). Regarding GHs targeting cellulose, xylan, and chitin, the GeneHunt approach provided a systematic and consistent annotation for potential cellulases from GH5, GH6, GH7, GH8, GH9, GH12, GH44, GH48, for potential xylanases from GH10 and GH11, and for potential chitinases from GH18, GH19, and GH85. Conversely in a few GH families including GH16, GH22, GH52, and potential cellulases from GH45 and xylanases from GH30, fewer than 90% of the proteins were annotated consistently using the GeneHunt approach. Among the GH families that could not be identified were some families that have been reclassified (e.g., GH61, GH69) and some GH families with reduced number of sequences. Having no specific HMM profile, members of these families were eventually assigned to other GH families (e.g., GH 74, 82, 84, 86). Finally, some GH families were associated with several Pfam IDs corresponding to various subdomains (e.g., N- and C-terminal domains) such as GH30, GH36, GH49, and GH79. Globally, the sequences from these GH families with questionable Pfam-based annotation accounted for <3% of the analyzed sequences.

Next, we used the GeneHunt approach to identify GH domains in the M5nr database and investigated the exact domain associations among MDGHs (Table S1, Supplementary Data 2). The GeneHunt approach identified 322,068 unique protein sequences with GH domains among the 43,098,145 non-redundant sequences (~0.7%). The most abundant domains were from the GH13 α-amylase (n = 47,737), GH34 neuraminidase (n = 22,357), GH3 β-glucosidase (n = 21,503), and GH1 glucosidase (n = 17,715) families. Conversely, in some GH families we identified a reduced number of proteins including 16, 378, and 389 chitosanases from GH80, GH75, and GH46, respectively.

All the domains existed as single domain protein (i.e., SDGH). More precisely, >90% of the identified proteins with a domain from GH families 1, 7, and 34 were SDGHs (Fig. 1A), whereas most domains from GH families 2, 4, and 30, among others, were identified in MDGHs (Fig. 1A). Regarding the domains targeting cellulose, xylan, and chitin (Fig. 1B), most GH7s and GH8s were SDGHs, whereas ~25% of the domains from GH families 5, 11, 12, and 45 were found in MDGHs. Next, potential cellulases from GH families 6, 9, 44, and 48, xylanases from GH family 10, and chitinases from GH families 18, 19, and 85 were more frequently found associated with other domains. Finally, 86% of the identified GH30 domains were found in multi-domain proteins (mostly, in association with the subdomain GH30c, Fig. 1B).

Associated with these GH domains, we found many potential non-GH CAZy domains including lipases (e.g., PF13472, PF00151) and polysaccharide deacetylases (e.g., PF01522) listed as carbohydrate esterase (CEs) in the CAZy database (Fig. 2A). We also identified many non-catalytic accessory CBMs and many other domains such as 6,672 F5/F8 type $C_{PF00754}$ domains associated with GH20, GH2, GH5, GH13, GH43, and GH30 and 1,013 $FIVAR_{PF07554}$ domains associated with GH85, GH20, GH31, GH43 and GH13 whereas 698 $BIG\_2_{PF02368}$ domains were associated with GH32, GH3, GH13, GH42, GH43, and GH10 among others.

Potential MDGHs targeting cellulose, xylan, and chitin consisted mostly of only 2 associated domains. However, more complex proteins with >8 domains were also identified (Fig. 1B). Domains from GH9, GH10, GH44, and GH85 were the most frequently identified domains in these complex multi-domain proteins. For example, 82 different protein domains were identified associated with 3,619 GH9s. However, 27 domains were observed only once (e.g., $Trypsin_{PF00089}$, $LPMO\_10_{PF03067}$, $B\text{-}lectin_{PF01453}$) whereas 401 $CBM3_{PF00942}$, 307 $Dockerin\text{-}1_{PF00404}$, 187 $CBM2_{PF00553}$, 1,305 $CelD\_N_{PF02927}$, and 89 $fibronectin\text{-}3_{PF00041}$ were the most abundant domains associated with GH9, sometimes in complex associations (Fig. 3A).

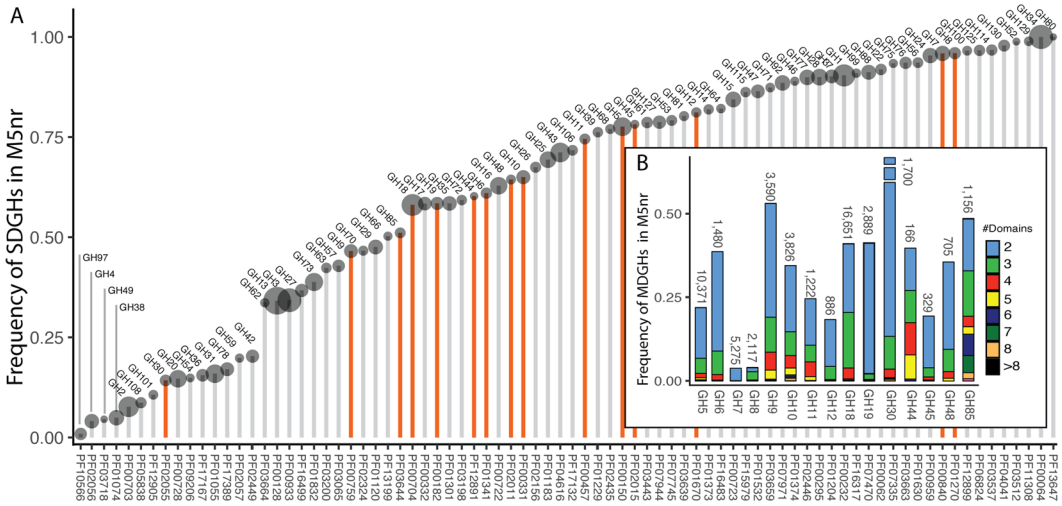

Figure 1. (**A**) Frequency of SDGHs identified in the M5nr database, size of dots mirrors the total number of identified domains (Supplementary Data). Highlighted in orange are the GH family detailed in (**B**). (**B**) Frequency of the complex MDGHs architecture in potential cellulases (i.e., GH5, 6, 7, 8, 9, 12, 44, 45, and 48), xylanases (i.e., GH10, 11, and 30), and chitinases (i.e., GH18, 19, and 85) identified in M5nr. Numbers correspond to the total number of identified domains in the M5nr database.

In order to further investigate the diversity of MDGHs with potential for cellulose, xylan, and chitin deconstruction we next investigated the domain-co-occurrence network and focused on associations observed more than once (Fig. 2A). Regarding potential multi-domain cellulolytic enzymes, domains from GH5, GH6, GH9, GH12, GH44, and GH48 formed a large cluster also containing potential xylanases whereas GH8 was clustered with potential chitinases. In the large cellulolytic cluster, we identified 52 MDGHs with more than one GH5 (i.e., multi-GH5), and few containing up to 4 GH5 domains (Fig. 3A). Next, the GH5 domain was found in many multi-domain associations with non-catalytic domains (e.g., $CBM2_{PF00553}$, $CBM5/12/2_{PF14600}$) and several catalytic domains including GH6 (n = 6), GH9 (n = 3), GH10 (n = 4), GH11 (n = 11), GH12 (n = 16), GH18 (n = 12), and GH44 (n = 5) (Figs 2A and 3A). The other cellulases within this large cellulolytic/xylanolytic cluster displayed similar types of associations (Supplementary Data). Interestingly, we identified many MDGHs with domain repetition including 519 multi-GH9s, 8 multi-GH6s, 16 multi-GH7s, and 10 multi-GH44s, whereas no multi-GH45 nor multi-GH48 were identified. Beside these associations, only 3 out of 5,290 identified GH7 were associated with $CBM1_{PF00734}$ whereas many GH45s (n = 339) were associated to $CBM2_{PF00553}$ or $CBM10_{PF02013}$. Finally, the 2,117 identified GH8, were found in none of the previously identified association, formed no multi-GH8, and clustered with potential chitinolytic domains (e.g., GH18).

Next, we identified two main clusters of multi-domain xylanases. The first one, with GH10 and GH11 clustered within the large cluster of previously identified potential cellulases. This cluster contained 30 multi-GH10s, 21 multi-GH11s, and 9 GH10-GH11 (Fig. 3A). We also identified several enzymes with non-catalytic accessory domains listed in the CAZy database such as CBMs (e.g., $CBM1_{PF00734}$ and $CBM4\_9_{PF02018}$) and $Dockerin_{PF00404}$ domains for cellulosome assembly (Fig. 3A). Several GH10s and GH11s were associated with other GH domains (e.g., GH5, GH12), or non-GH CAZyme domains such as Polysaccharide $Deacetylase_{PF01522}$. Finally, several other domains, such as $Esterase_{PF00756}$, were associated to GH10 and GH11 (Fig. 3A). The GH30, and its associated subdomain GH30c, formed a distinct and large cluster containing 56 multi-GH30s and displaying many domain associations not found with other xylanase domains. These included associations with other GH domains (e.g., GH3, GH13, GH43), non-GH CAZyme domains (e.g., $CBM4/9_{PF02018}$, $CBM6_{PF03422}$) and many other domains such Lipase $GDSL\text{-}2_{PF1347}$, Ricin B lectin 2_{PF1420}, and several domains with unknown function (e.g., $DUF5011_{PF16403}$, Figs 2A and 3A).

Regarding potential chitinases, we identified 277 multi-GH18s and 19 multi-GH19s. In addition, GH18 and GH19 were associated in one GH18-GH19 (Fig. 3A). Next, chitinases were associated with many of the previously listed GH domains (e.g., 12 GH18-GH5), and other GH domains such as $LysM_{PF01476}$ (GH25, n = 4). In addition, several of these potential chitinases contained CBMs such as $CBM14_{PF01607}$ (n = 254), $CBM5/12_{PF02839}$ (n = 10), and $CBM2_{PF00553}$ (n = 32). Finally, among the 1,163 listed GH85 domains, 7 multi-GH85s were identified. We also identified 221 $Big\text{-}3_{PF07523}$, 227 $F5\_F8\_type\_C_{PF00754}$, and 176 $FIVAR_{PF07554}$ associated to GH85, among others (Fig. 3A).

Mapping of GHs in assembled metagenomes. Next, we used the GeneHunt approach to identify GH domains in 129 publicly accessible assembled metagenomes from MG-RAST (Table S1, Supplementary Data 3) and identified 200,257 GH domains corresponding to 160,790 proteins. Across datasets, potential α-amylases from GH13 were the most abundant domains (n = 28,135) followed by potential β-glucosidase from GH3 (n = 18,403), β-xylosidase/α-L-arabinofuranosidase from GH43 (n = 11,609) and β-galactosidase/β-mannosidase from GH2 (n = 10,571). As described for MDGHs in the M5nr database, we identified many domains associated to GHs

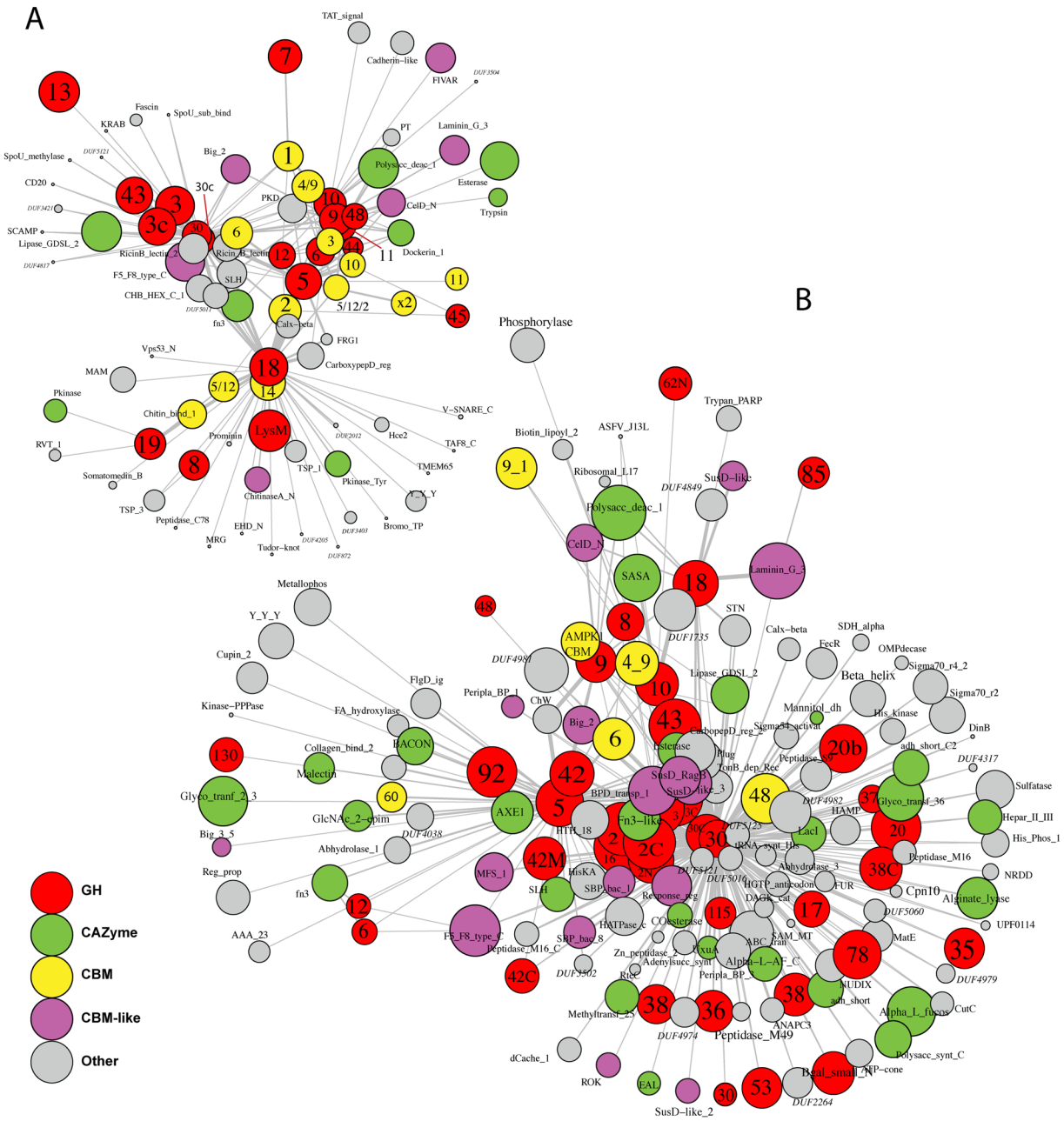

Figure 2. Association map for multi-domain cellulolytic, xylanolytic, and chitinolytic domains in the M5nr database (**A**) and in 129 combined assembled metagenomes (**B**), generated using igraph.

including several well-known CBMs (e.g., CBM2), some non-GH CAZymes (e.g., glycosyl-transferase$_{PF00535}$), and many catalytic (e.g., phosporylase$_{PF00343}$, peptidase_M16$_{PF00675}$) and non-catalytic domains (e.g.,DUF4979$_{PF16351}$, Calx-beta$_{PF03160}$) not listed on the CAZy database.

Next, potential domains for cellulose deconstruction were dominated by 5,004 GH5 domains distributed in 4,921 proteins and 82 multiGH5s. Next, we identified 1,459 GH9s, 85 GH6s (1 multiGH6), 9 GH7s, 746 GH8s (1 multiGH8), 71 GH12s, 132 GH44s, 17 GH45s (1 multiGH45), and 36 GH48s (4 multiGH48s). Noteworthy, beside potential cellulases with repeated domains, 97 MDGH cellulases displayed at least 2 potential catalytic domains targeting cellulose (e.g., GH9-GH8) (Figs 3B and 4B). Xylanase domains consisted in 2,175 GH10s (99 multiGH10s), 43 GH11s (1 multiGH11), and 1,547 GH30s (136 multiGH30s). Globally, 3,524 potential xylanases were identified; 3,288 proteins with only one xylanase domain, 2,224 being SDGH proteins, and 236 proteins with at least two potential domains for potential xylanase. Conversely, 1,300 potential xylanases were MDGHs.

Finally, chitinase domains including 3,286 GH18s (25 multiGH18s), 627 GH19s, and 247 GH85s (1 mGH85) corresponded to 3,888 proteins including 2,638 SDGHs and 1,250 MDGHs. Interestingly, 26 multi-domain chitinases displayed 2 potential chitinolytic domain.

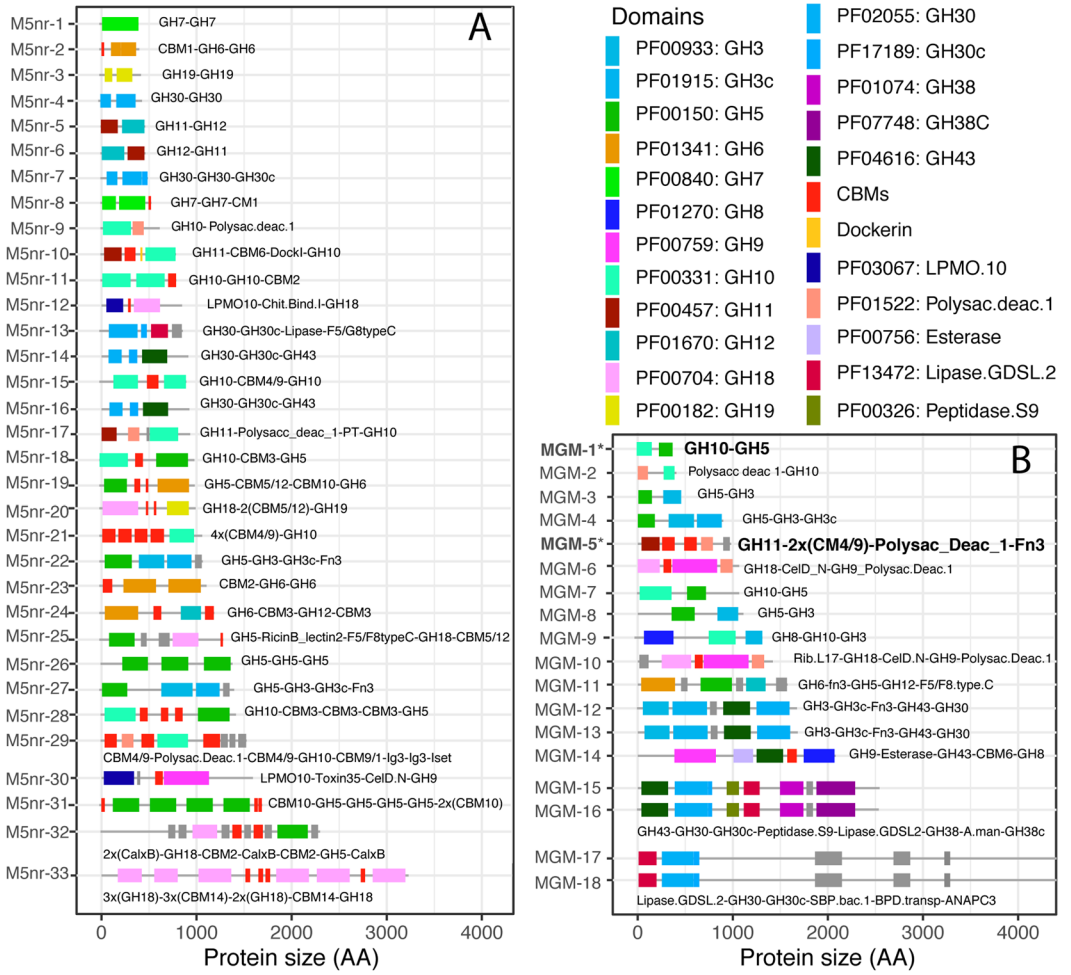

Figure 3. GeneHunt-based identification of new potential Multi-Activity GHs in the M5nr database (**A**) and in publicly accessible assembled metagenomes from MG-RAST (**B**). All identified proteins from MG-RAST (**B**) are from distinct metagenomes (Supplementary Data). CBMs (in red) include CBM1, 2, 3, 4/9, 5/12, 5/12-2, 6, 9-1, 10, 14, and chitin binding domain 1. MGM-1* and MGM-5* from assembled metagenomes were selected for further characterization (see text).

Beside these multi-domain cellulases, xylanases, and chitinases we identified several potential cellulase-xylanase (e.g., GH5-GH11), cellulase-chitinase (e.g., GH5-GH18), xylanase-chitinase (e.g, GH11-GH18), and several other potential multi-activity enzymes (Fig. 3B, Table S3). Although many domain combinations were unique, some have been identified multiple times.

Characterization of multi activity GHs. Overall, of the ~483,000 proteins sequences with at least one GH domains identified here, many contained non-catalytic accessory domain(s) and a few contained several catalytic domains. These corresponded to potential multi-activity proteins (i.e., MAGHs) with repeated domains such as proteins with multiple GH18 or multiple GH5 domains or proteins with distinct catalytic domains combined together (Fig. 3, Table S3). Some identified MAGHs were unique whereas some, even very complex, were identified multiple times and in different datasets. For example, a few very long and complex proteins with up to 7 distinct domains including 6 different potential catalytic domains were identified in several metagenomes (Fig. 3B). Although rare, being identified in multiple and unrelated datasets supports the biological origin of these complex proteins rather than *in silico* artifacts.

Based on the domain associations and knowing the function of several GH families, some identified MAGHs potentially displayed interesting synergies among the catalytic domains. Thus, in order to further demonstrate the biotechnological potential of these MAGHs we selected two unique proteins identified in distinct environmental datasets for sequence optimization and gene synthesis to proceed with heterologous protein production in *E. coli*.

First, MGM-1 (i.e., mgm4441594_JCVI_READ_1095454020156_1_1140_-) is a 377 amino acid protein, identified in a marine metagenome[36], consisting of a potential cellulase from GH5 associated with a potential xylanase from GH10 (Fig. 3B). As expected, recombinant MGM-1 produced in *E. coli* was active on both AZCL-xylan and CMC (Fig. 4A,B).

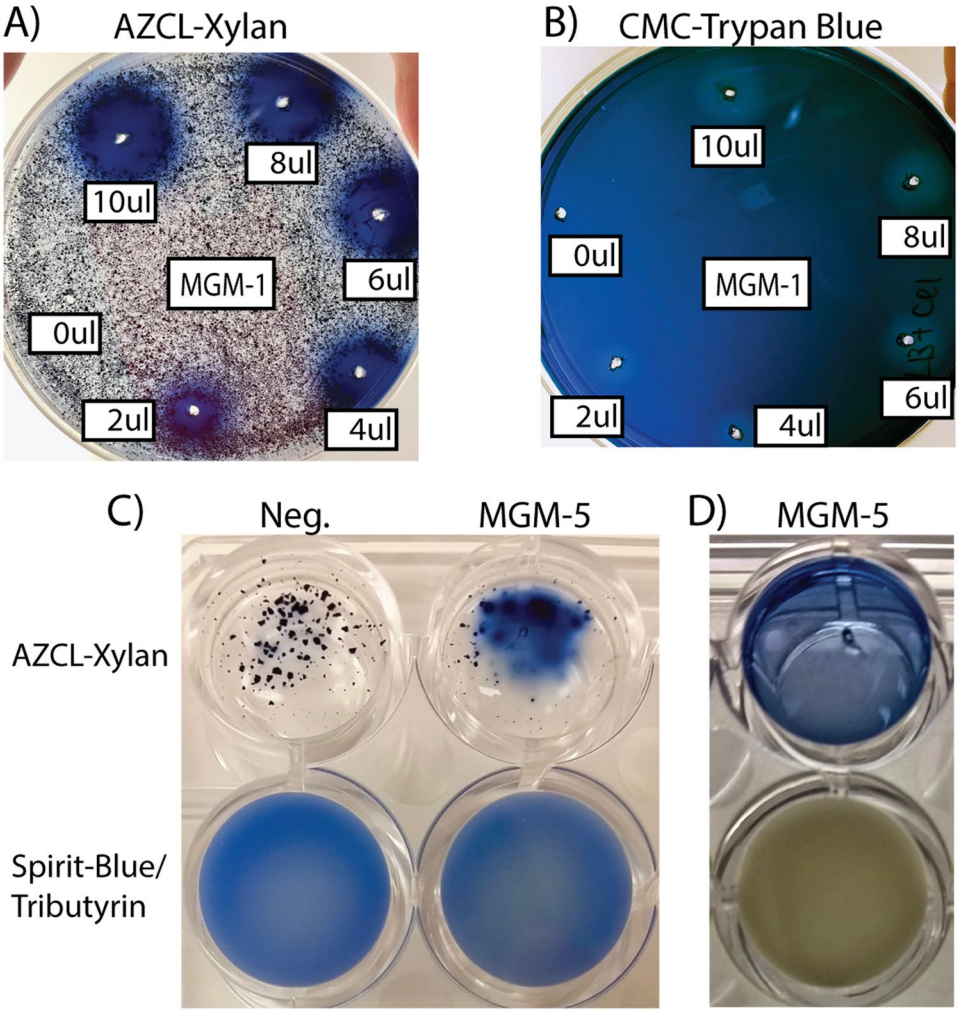

Figure 4. Biochemical activities of recombinant enzymes MGM1 (GH10-GH5) and MGM-5 (GH11-2 × CBM4/9-Polysacc.Deac.1-Fn3) (see Fig. 3). Before incubation, 0 to 10 µl of the MGM-1 extract in 20 mM Tris-HCl (pH 8.0) were mixed with 10 to 0 µl of 20 mM Tris-HCl (pH 8.0) to test the effect of protein concentration on AZCL-Xylan (**A**) and CMC-Trypan Blue (**B**). For MGM-5, 20 µl of cell extract in 20 mM Tris-HCl (pH 8.0) was incubated for 4 h (**C**) and 48 h (**D**) with AZCL-Xylan and Tributyrin. (Negative control was the cell extract of non-induced E. coli BL21(DE3):pet151B-MGM-5).

Next, MGM-5 with 879 amino acids (i.e., mgm4491477_NODE_11875_length_6674_cov_6.617920_2824_5466_-), derived from a human gut sample[37], contained a potential GH10-xylanase associated with a domain identified as a potential polysaccharide deacetylase and 2 CBM4/9s$_{PF02018}$ (Fig. 3B, Table S3). As expected, recombinant MGM-5 hydrolyzed both AZCL-xylan and tributyrin (Fig. 4C,D).

Discussion

The GeneHunt approach, using publicly accessible HMM-profiles from the Pfam A database[29], can be used to identify the vast majority of characterized GHs (>98%) listed on the CAZy database[1]. The identification of rare or newly defined GH families such as the GH family 156 (introduced in October 2018)[1] with just 7 identified sequences, having no HMM-profile available, is not yet possible using this approach[30]. For these families, until more sequences are identified and HMM-profiles created, similarity searches (e.g., BLAST[38]) using a custom database is an alternative (see Supplementary Data 1). However, as described for bacteria[2] and fungal genomes[10], the GeneHunt approach can identify the detailed architecture of most GH families in large database and in metagenomes. Instead of using a single-step HMMscan[28] with a custom database (e.g., dbCAN[39]), the GeneHunt approach performs two sequential HMM-profile identifications. The first search, for all the protein sequences, is performed using a small custom HMM-profile database whereas the second scan, only for the potential positive hits derived from the first step, uses the entire Pfam A database[29]. This approach identifies all the domains associated with GHs including the ones not listed in the custom database while minimizing the number of sequence analyses, and thus is faster than a direct scan against the complete Pfam A database. This approach can be adjusted at will by searching new HMM-profiles in the small custom database, including HMM-profiles derived from dbCAN and other domains of interest such a susD-transporters$_{PF12741}$ or lipases$_{PF00151}$. In this context, the GeneHunt approach

allows the identification and investigation of GH architecture in large databases such as sequenced bacterial genomes from the PATRIC database[2,40], fungal genome from the Mycocosm database[10,41] and MG-RAST[32], as described here.

In addition to identifying 483,000 sequences for GHs in a curated database (CAZy database[1]), in a large non-redundant database (M5nr), and in 129 environmental datasets from MG-RAST[32], we identified hundreds of accessory non-catalytic and catalytic domains associated with GH domains. Most identified MDGHs consisted of a GH domain associated with some non-catalytic accessory domain (e.g., CBM). Beside many well-characterized domains for non-GH CAZymes and carbohydrate binding modules (CBM)[1], hundreds of domains are associated with GHs. Among others, the $FIVAR_{PF07554}$ domain, found in various architectures, binds fibronectin and is sometimes linked to methicillin resistance[42,43], whereas the F5/F8 type $C_{PF00754}$ binds phospholipid on the surface of endothelial cells and adheres to glycoprotein in bovine milk[44,45]. The BIG_2 domain has been shown to be involved in cell-adhesion[46]. The systematic association of these various domains with GH domains provides insights on the modular nature of GHs in microbes. It has been shown that GHs can have multiple non-catalytic domains such as SusD-like $domains_{PF12741}$[47] or $CelD\_N_{PF02927}$ domains[48], which mediate xyloglucan-binding for cellular intake or cellulose binding by certain cellulases respectively. These non-catalytic domains play a major role in substrate binding, and thus potentially, affect the overall catalytic efficiency of associated catalytic domains. The frequent association of poorly characterized domains (e.g. domains of unknown function - DUF) could be used to infer and test potential domain activities. Based on their frequency one could identify poorly characterized domains (e.g., DUF4979) systematically associated with specific GH domains to infer and test their function.

Additionally, a few MAGHs, displayed several catalytic domains and thus potentially combine distinct enzymatic activities. The vast majority of characterized GHs, including the ones listed on the CAZy database, targeting structural polysaccharides are single domain microbial GHs displaying low to moderate activity on natural substrates[30]. This highlights the need for multiple catalysts acting synergistically to support the efficient biomass deconstruction[20,49]. Linked multi-activity complexes such as cellulosomes[50] and MAGHs[33,35,51] display increased synergistic interaction amongst domains and represent an interesting alternative to complex mixtures of enzymes. Indeed, in cellulosomes and MAGHs, beside the additive effect of the catalytic domains there exist a proximity effect that reduces the diffusion of the catalytic domains relative to each other. However, although limited in the number of associated domains, MAGHs have the advantage of being stable covalent complexes, unlike cellulosome[52] and the few characterized MAGHs display high hydrolytic activity[21,33,35,51].

Identifying the complete set of domain combinations in MDGHs and MAGHs is a prerequisite to investigate the evolution of the protein domains from simple to complex multi-domain enzyme[53]. In addition, because the functions of many GHs families are conserved[1,30], it is possible to infer how the combined domains could interact. Different types of synergy can be envisioned in MAGHs including linear pathway synergy (LPS), parallel pathway synergy (PPS), and debranching synergy (DS). In LPS the first catalytic domain is expected to release the substrate of the second catalytic domain whereas in PPS (e.g., MGM-1) the catalytic domains target distinct yet physically associated substrates. Finally, in DS the first catalytic domain cleaves the side groups in substituted polysaccharide thus increasing the accessibility of the polysaccharide backbone (e.g., MGM-5). Although rare, these MAGHs with multiple catalytic domains represent potential robust hydrolytic systems with reduced inhibition by the product[50,54] and display proximity effect analogous to carbohydrate binding modules[3]. In addition, MDGHs with DS and including some esterase activity, could possibly disrupt the xylan-lignin complex and thus improve the xylan deconstruction by associated xylanases[55,56]. Finally, depending on the processivity (the enzyme's ability to catalyze several consecutive reactions without releasing the substrate) of individual catalytic domain, some MAGHs could display unique modes of action[21,35].

Finally, MAGHs combining distinct catalytic domains, while being encoded by single genes, can easily be edited (e.g., tagging the protein), cloned, and expressed in various hosts. In this context, the ever-growing number of accessible sequences-datasets (i.e., genomes and metagenomes) provides an unprecedented opportunity to identify new biotechnologically interesting catalysts. In addition, investigating the domain association in nature also highlights new ways to associate protein domains in order to take advantage of nature diversity for the purpose of synthetic biology.

Methods

GeneHunt approach. Briefly, the GeneHunt approach provides a Pfam-based domain-specific annotation of protein sequences[2]. More precisely, GeneHunt uses sequential HMM-profile searches[28] to rapidly identify the detailed multi-domain organization of proteins containing a domain of interest (e.g., PF00150 for GH5). First, selected HMM-profiles for domains of interest (Table S1) derived from the Pfam-A database[29] are searched (HMMsearch) in protein datasets. Then, the protein sequences of the potential positive hits are scanned against the entire Pfam-A database (HMMscan). The first search is fast and inaccurate, whereas the second scan identifies all the domains, not just the domains of interest, and removes the false positive hits. Although relatively slow, this second scan is performed on narrowed sets of sequences, thus making the overall process faster than a direct comparison of the entire dataset using the entire Pfam-A database. GeneHunt is publicly accessible on https://github.com/renober/GeneHunt_V1.

Datasets. To test the GeneHunt approach, we first manually retrieved and reannotated 7,920 sequences for biochemically characterized GHs listed on the CAZy database[1], as of June 2018 (Supplementary Data). Next, we retrieved and reannotated the M5nr database (ftp.metagenomics.anl.gov/, version 2013.12.15) containing $\sim 43 \times 10^6$ mostly microbial non-redundant protein sequences[31]. Finally, we retrieved the protein sequences from 129 publicly accessible assembled metagenomes from MG-RAST (Table S2) using the MG-RAST's "application programming interface" (API)[57].

DNA synthesis and protein expression. The DNA sequences of two potential multi-activity GHs were first optimized for expression in *E. coli*, and cloned in the pET151 in order to incorporate the pelB signal peptide in the N-terminal end and a His-tag at the C-terminal end of the proteins (Thermo Fisher, Vista, CA., USA). The plasmids were then introduced into competent *E. coli* BL21(DE3) (Novagen, Madison, WI, USA). Heterologous protein expression was carried out in Lysogenic Broth at 37 °C for four hours by adding 0.4 mM isopropyl-D-1-thiogalactopyranoside (isopropyl-beta-thio-galactoside, IPTG) when the OD_{600nm} reached ~0.5. After centrifugation, the cell pellet was resuspended in 20 mM Tris-HCl (pH 8.0) and the cells were disrupted by sonication. Proteins from the cytoplasmic fraction were recovered by centrifugation at 20,000 × g for 40 min. Then enzymatic activities were tested qualitatively using chromogenic substrates. More precisely, azurin-cross linked xylan (AZCL-Xylan, Megazyme, Chicago, IL., USA) was used to detect xylanase, Trypan-Blue:CarboxyMethyl-Cellulose (CMC, Sigma-Aldrich, St Louis, MO., USA)[23] was used for cellulase, and esterase activity was tested by incubating the extract with tributyrin in presence of the pH-indicator Spirit Blue[58]. Xylanolytic and cellulolytic activities were visualized after incubation for 4 hours at room temperature whereas tributyrin hydrolysis required a 48 hours incubation. Thus, in order to discriminate the recombinant activity from "residual activity" from the *E. coli* BL21(DE3) used for protein production, the cytoplasmic fraction of non-induced cells was used as negative control for tributyrin assay.

Data processing and availability. Data were processed using R (Version 1.1.456) and the packages ggplot2, gplots, plyr, dplyr, reshape, reshape2, and igraph. PFam-based annotation of biochemically characterized GH listed on the CAZy database are in Supplementary File 1 and include the sequence ID, the taxonomic origin, the original annotation from CAZy, the EC-classification, and the GH domains identified using GeneHunt. Detailed GH-sequences annotation derived from M5nr database is in Supplementary File 2. Sequence from the M5nr database can be retrieved directly from the MG-RAST portal using MG-RAST API[57] synchronous GET requests: http://api.metagenomics.anl.gov/m5nr/md5/SequenceID?sequence=TRUE).

Detailed GH-sequences annotation derived from M5nr database is in Supplementary Data 3. Sequences can be retrieved using MG-RAST API[57] synchronous GET requests: http://api.metagenomics.anl.gov/download/mgmid?stage=650.

Data Availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information Files). In addition, GeneHunt_V1.sh is publicly available on GitHub (https://github.com/renober/GeneHunt_V1).

References

1. Lombard, V. *et al*. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–5 (2014).
2. Talamantes, D., Biabini, N., Dang, H., Abdoun, K. & Berlemont, R. Natural diversity of cellulases, xylanases, and chitinases in bacteria. *Biotechnol. Biofuels* **9**, 133 (2016).
3. Hervé, C. *et al*. Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proc. Natl. Acad. Sci. USA* **107**, 15293–8 (2010).
4. Várnai, A., Siika-Aho, M. & Viikari, L. Carbohydrate-binding modules (CBMs) revisited: reduced amount of water counterbalances the need for CBMs. *Biotechnol. Biofuels* **6**, 30 (2013).
5. Allison, S. D. *et al*. Microbial abundance and composition influence litter decomposition response to environmental change. *Ecology* **94**, 714–25 (2013).
6. Baldrian, P. *et al*. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *ISME J.* **6**, 248–58 (2012).
7. Souza, C. P., Almeida, B. C., Colwell, R. R. & Rivera, I. N. G. The importance of chitin in the marine environment. *Mar. Biotechnol. (NY).* **13**, 823–30 (2011).
8. Gutowska, M. A., Drazen, J. C. & Robison, B. H. Digestive chitinolytic activity in marine fishes of Monterey Bay, California. *Comp. Biochem. Physiol. Part A Mol. Integr. Physiol.* **139**, 351–358 (2004).
9. Lindahl, B. D. & Finlay, R. D. Activities of chitinolytic enzymes during primary and secondary colonization of wood by basidiomycetous fungi. *New Phytol.* **169**, 389–397 (2006).
10. Berlemont, R. Distribution and diversity of enzymes for polysaccharide degradation in fungi. *Sci. Rep.* **7**, 222 (2017).
11. Treseder, K. K. & Lennon, J. T. Fungal traits that drive ecosystem dynamics on Land. *Microbiol. Mol. Biol. Rev.* **79**, 243–262 (2015).
12. Berlemont, R. & Martiny, A. C. Genomic potential for polysaccharides deconstruction in bacteria. *Appl. Environ. Microbiol.* **81**, 1513–19 (2015).
13. Tauzin, A. S. *et al*. Molecular Dissection of Xyloglucan Recognition in a Prominent Human Gut Symbiont. *MBio* **7**, e02134–15 (2016).
14. Tamura, K. *et al*. Molecular Mechanism by which Prominent Human Gut Bacteroidetes Utilize Mixed-Linkage Beta-Glucans, Major Health-Promoting Cereal Polysaccharides. *Cell Rep.* **21**, 417–430 (2017).
15. Howe, A., Yang, F., Williams, R. J., Meyer, F. & Hofmockel, K. S. Identification of the Core Set of Carbon-Associated Genes in a Bioenergy Grassland Soil. *PLoS One* **11**, e0166578 (2016).
16. Warnecke, F. *et al*. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560–5 (2007).
17. Knight, R. *et al*. Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* **30**, 513–20 (2012).
18. Hess, M. *et al*. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–7 (2011).
19. Himmel, M. E. & Bayer, E. A. Lignocellulose conversion to biofuels: current challenges, global perspectives Editorial overview. 316–317, https://doi.org/10.1016/j.copbio.2009.05.005 (2009).
20. Dodd, D. & Cann, I. K. O. Enzymatic deconstruction of xylan for biofuel production. *Glob. Change Biol. Bioenergy* **1**, 2–17 (2009).
21. Brunecky, R. *et al*. The Multi Domain Caldicellulosiruptor bescii CelA Cellulase Excels at the Hydrolysis of Crystalline Cellulose. *Sci. Rep.* **7**, 9622 (2017).
22. Sathya, T. A. & Khan, M. Diversity of Glycosyl Hydrolase Enzymes from Metagenome and Their Application in Food Industry. *Concise Rev. Food Sci.* **79**, 2149–2156 (2014).
23. Berlemont, R. *et al*. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. *ISME J.* **3**, 1070–1081 (2009).

24. Nyyssönen, M. *et al*. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Front. Microbiol.* **4**, 282 (2013).
25. Berlemont, R. & Martiny, A. C. Glycoside Hydrolases across Environmental Microbial Communities. *PLOS Comput. Biol.* **12**, e1005300 (2016).
26. Muegge, B. D. *et al*. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**, 970–4 (2011).
27. King, A. J. *et al*. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proc. Natl. Acad. Sci.* **107**, 5345–5350 (2010).
28. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
29. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–30 (2014).
30. Nguyen, S. T. C., Freund, H. L., Kasanjian, J. & Berlemont, R. Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy. *Appl. Microbiol. Biotechnol.* **102**, 1629–1637 (2018).
31. Wilke, A. *et al*. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* **13**, 141 (2012).
32. Keegan, K. P., Glass, E. M. & Meyer, F. In 207–233, https://doi.org/10.1007/978-1-4939-3369-3_13 (2016).
33. Zhang, C. *et al*. Characterization of a multi-function processive endoglucanase CHU_2103 from Cytophaga hutchinsonii. *Appl. Microbiol. Biotechnol.* **98**, 6679–6687 (2014).
34. Kim, S.-K., Chung, D., Himmel, M. E., Bomble, Y. J. & Westpheling, J. *In vivo* synergistic activity of a CAZyme cassette from *Acidothermus cellulolyticus* significantly improves the cellulolytic activity of the *C. bescii* exoproteome. *Biotechnol. Bioeng.*, https://doi.org/10.1002/bit.26366 (2017).
35. Brunecky, R. *et al*. Revealing nature's cellulase diversity: the digestion mechanism of Caldicellulosiruptor bescii CelA. *Science* **342**, 1513–6 (2013).
36. Rusch, D. B. *et al*. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
37. Claesson, M. J. *et al*. Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178–184 (2012).
38. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).
39. Yin, Y. *et al*. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).
40. Wattam, A. R. *et al*. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* **45**, D535–D542 (2017).
41. Grigoriev, I. V. *et al*. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–704 (2014).
42. Williams, R. J., Henderson, B., Sharp, L. J. & Nair, S. P. Identification of a Fibronectin-Binding Protein from Staphylococcus epidermidis. *Infect. Immun.* **70**, 6805–6810 (2002).
43. Komatsuzawa, H., Ohta, K., Sugai, M., Fujiwara, T. & Glanzmann, P. JAC Tn 551 -mediated insertional inactivation of the fmtB gene encoding a Staphylococcus aureus. *J. Antimicrob. Chemother.* **45**, 421–431 (2000).
44. Veeraraghavan, S., Baleja, J. D. & Gilbert, G. E. In the presence of dodecylphosphocholine micelles. *J. Biochem.* **332**, 549–555 (1998).
45. Hvarregaard, J., Andersen, M. H., Berglund, L., Rasmussen, J. T. & Petersen, T. E. Characterization of glycoprotein PAS-6/7 from membranes of bovine milk fat globules. *Eur. J. Biochem.* **240**, 628–636 (1996).
46. Kelly, G. *et al*. Structure of the cell-adhesion fragment of intimin from enteropathogenic Escherichia coli. *Nat. Am. Inc.* **6**, 313–318 (1999).
47. Larsbrink, J. *et al*. A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
48. Dominguez, R., Lascombe, M., Alzari, P. M. & Couchon, H. The Crystal Structure of a Family 5 Endoglucanase Mutant in Complexed and Uncomplexed Forms Reveals an Induced Fit Activation Mechanism. *J. Mol. Biol.* **257**, 1042–1051 (1996).
49. Wilson, D. B. Microbial diversity of cellulose hydrolysis. *Curr. Opin. Microbiol.* **14**, 259–63 (2011).
50. Gefen, G., Anbar, M., Morag, E., Lamed, R. & Bayer, E. A. Enhanced cellulose degradation by targeted integration of a cohesin-fused β-glucosidase into the Clostridium thermocellum cellulosome. *Proc. Natl. Acad. Sci. USA* **109**, 10298–303 (2012).
51. Gibbs, M. D. *et al*. Multidomain and multifunctional glycosyl hydrolases from the extreme thermophile Caldicellulosiruptor isolate Tok7B.1. *Curr. Microbiol.* **40**, 333–40 (2000).
52. Smith, S. P. & Bayer, E. A. Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. *Curr. Opin. Struct. Biol.* **23**, 686–694 (2013).
53. Forslund, K. & Sonnhammer, E. L. L. Evolution of protein domain architectures. *Methods Mol. Biol.* **856**, 187–216 (2012).
54. Prawitwong, P. *et al*. Direct glucose production from lignocellulose using Clostridium thermocellum cultures supplemented with a thermostable β-glucosidase. *Biotechnol. Biofuels* **6**, 184 (2013).
55. Dilokpimol, A. *et al*. Fungal glucuronoyl esterases: Genome mining based enzyme discovery and biochemical characterization. *N. Biotechnol.* **40**, 282–287 (2018).
56. Dodd, D. *et al*. Biochemical analysis of a β-D-xylosidase and a bifunctional xylanase-ferulic acid esterase from a xylanolytic gene cluster in Prevotella ruminicola 23. *J. Bacteriol.* **191**, 3328–3338 (2009).
57. Wilke, A. *et al*. A RESTful API for accessing microbial community data for MG-RAST. *PLoS Comput. Biol.* **11**, e1004008 (2015).
58. Berlemont, R. *et al*. Novel Cold-Adapted Esterase MHlip from an Antarctic Soil Metagenome. *Biology (Basel).* **2**, 177–88 (2013).

Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R25GM071638 and 8UL1GM118979-02 (RB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

Conceptualization: S.N. and R.B. Data Curation and Analysis: S.N., D.T., F.D., A.V. and R.B. Molecular Biology: A.F., J.S. and R.B. Supervision: J.S. and R.B. Writing: S.N. and R.B.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-46290-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.