A subfamily roadmap for functional glycogenomics of the evolutionarily diverse Glycoside Hydrolase Family 16 (GH16)

Alexander Holm Viborg, Nicolas Terrapon, Vincent Lombard, Gurvan Michel, Mirjam Czjzek, Bernard Henrissat, Harry Brumer

A subfamily roadmap for functional glycogenomics of the evolutionarily diverse Glycoside Hydrolase Family 16 (GH16)

Alexander Holm Viborg, Nicolas Terrapon, Vincent Lombard, Gurvan Michel, Mirjam Czjzek, Bernard Henrissat, Harry Brumer

A subfamily roadmap for functional glycogenomics of the evolutionarily diverse Glycoside Hydrolase Family 16 (GH16)

Alexander Holm Viborg[1,=], Nicolas Terrapon[2,3,=], Vincent Lombard[2,3], Gurvan Michel[4], Mirjam Czjzek[4], Bernard Henrissat[2,3,*], and Harry Brumer[1,5,6,7,*]

From the [1]Michael Smith Laboratories, University of British Columbia, 2185 East Mall, Vancouver, BC V6T 1Z4, Canada; [2]Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, F-13288 Marseille, France; [3]USC1408 Architecture et Fonction des Macromolécules Biologiques, Institut National de la Recherche Agronomique, F-13288 Marseille, Universités, UPMC Univ Paris 06, CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, Roscoff, France; [5]Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, BC V6T 1Z1 Canada; [6]Department of Biochemistry and Molecular Biology, University of British Columbia, 2350 Health Sciences Mall, Vancouver, BC V6T 1Z3 Canada; [7]Department of Botany, University of British Columbia, 6270 University Blvd., Vancouver, BC V6T 1Z4 Canada.

Running title: *GH16 subfamilies*

[=] Equal contribution.

*To whom correspondence should be addressed. Harry Brumer: Michael Smith Laboratories, University of British Columbia, 2185 East Mall, Vancouver BC, Canada, V6T 1Z4, brumer@msl.ubc.ca, Tel. +1 604 827 3738; Bernard Henirssat: AFMB UMR 7257 Case 932, Campus de Luminy, 163 Avenue de Luminy, 13288 Marseille CEDEX 09; bernard.henrissat@afmb.univ-mrs.fr; Tel. +33 491 82 55 87

Keywords: phylogenetics, enzyme structure, protein evolution, structural biology, glycoside hydrolase, carbohydrate-active enzymes (CAZymes), sequence similarity networks (SSN), Hidden Markov Model (HMM), beta-jelly-roll fold, beta-sandwich

ABSTRACT

Glycoside Hydrolase Family 16 (GH16) comprises a large family of glycosidases and transglycosidases based on a common beta-jelly-roll fold, whose taxonomically diverse members are active on a range of terrestrial and marine polysaccharides. Presently, facile sequence-function correlations in GH16 are hindered by a lack of a systematic subfamily structure. Using a highly scalable protein Sequence Similarity Network (SSN) analysis, we have delineated nearly 23,000 GH16 sequences into 23 robust subfamilies, which are strongly supported by Hidden Markov Model (HMM) and Maximum Likelihood (ML) molecular phylogenetic analyses. Subsequent evaluation of over 40 experimental three-dimensional structures has highlighted key tertiary structural differences that dictate substrate specificity across the GH16 evolutionary landscape. As for other large GH families (*i.e.* GH5, GH13, and GH43), this new subfamily classification provides a roadmap for functional glycogenomics that will guide future bioinformatics and experimental structure-function analyses. The GH16 subfamily classification is publicly available in the CAZy database via URL www.cazy.org/GH16.html. The SSN workflow used here is available via URL https://github.com/ahvdk/SSNpipe/.

INTRODUCTION

Complex carbohydrates – oligosaccharides and polysaccharides of diverse residue and linkage composition – are central to a wide range of biological processes, such as energy storage, inflammation, host-pathogen interactions, diseases, and differentiation/development (1). Not least, manifold complex carbohydrates play essential structural roles in the cell walls in terrestrial and marine biomass (2, 3).These biomass sources represent major sinks in the global carbon cycle (4, 5) and a vast renewable resource for the production of energy, chemicals, and materials (6).

The synthesis, rearrangement, and ultimate saccharification of the vast diversity of glycosidic linkages in natural carbohydrates require a correspondingly broad range of specific carbohydrate-active enzymes (CAZymes). In light of the continually accelerating rate of sequence data deposition, the CAZy database has emerged as a central resource uniting specificity, mechanistic, and structural information within actively curated, sequence-based families of glycosyltransferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), auxiliary activity enzymes (AAs), and associated non-catalytic carbohydrate-binding modules (CBMs) (7, 8). The CAZy classification offers extraordinary predictive power on the family level, whereby the key active-site residues, the catalytic mechanism, and the overall three-dimensional fold are generally strictly conserved. Family classification is also a broad predictor of substrate specificity, in terms of overall glycosidic linkage orientation (alpha or beta) and saccharide composition. However, the subtle natural variations in configuration among structurally related groups of complex carbohydrates has given rise to several "polyspecific" families, which comprise diverse activities. As it pertains to genomics and bioinformatics, polyspecificity confounds precise functional annotation of CAZyme family members in the absence of biochemical data (7).

The problem of polyspecificity is especially significant among large CAZyme families, which may encompass tens-of-thousands of sequences from taxonomically diverse organisms. In such cases, division into subfamilies based on molecular phylogeny has been shown to significantly increase predictive power in a handful of GH and PL families previously (9–13). However, a major limitation of large-scale phylogenetic analyses is the dependency on a highly accurate Multiple Sequence Alignment (MSA) (14) and subsequent phylogenetic tree estimation, in which the computational complexity increases exponentially with the number of sequences (15). As the number of non-redundant sequences in the CAZy database increases (7), highly accurate subfamily phylogenies will be infeasible for most families in the foreseeable future.

Sequence similarity networks (SSNs), which are conceptually illustrated in Figure 1, offer a potential solution to this conundrum. In contrast to MSA-based phylogenies, SSNs are based on all-versus-all pairwise local sequence alignments, the computational requirements of which scales

linearly with the number of sequences and are easily amenable to parallelization. Notably, the resulting networks of nodes and edges, which can be rapidly generated using any Expect (*E*) value or bit-score as a threshold, usually resolve the same monophyletic groups observed in corresponding phylogenetic trees (16). Like phylogenetic approaches, SSNs can underpin the creation of subfamilies and establish a robust framework to predict substrate specificity and highlight unexplored sequence space (17).

Glycoside Hydrolase Family 16 (GH16) is a polyspecific family of β-glycanases involved in the degradation or remodeling of cell wall polysaccharides in marine and terrestrial biomass (Table 1). GH16 represents a current challenge for functional subfamily classification due to its large size and diversity. GH16 members are widely distributed across the domains of life, including bacteria (18), oomycetes (19), fungi (20, 21), plants (22, 23), and animals (terrestrial insects and marine invertebrates (24, 25)), in which they play manifold biological roles. GH16 members are united by a compact (*ca.* 30 kDa) β-jelly roll structural fold (26), which nonetheless has a remarkable evolutionary plasticity that gives rise to specificities for a plethora of complex terrestrial and marine cell-wall carbohydrates, hydrolase and transglycosylase activities, and non-catalytic substrate-binding functions (21, 27–29). Presently, GH16 comprises *ca.* 8000 sequences in the public CAZy database representing 15 known activities (7), which is comparable to other large families (GH5, GH43) for which subfamily classifications have been established (GH13 is an exception, with nearly 10-fold more members, while GH30 is four-fold smaller than GH16) (9–13). Only 2.5% of GH16 sequences have been enzymatically characterized (7), which challenges functional prediction.

Here we present a comprehensive subfamily classification of GH16 based on large-scale SSN analysis of the entire GH16 sequence space as a roadmap for future functional glycogenomics. The subfamily topology was equal to that obtained by classical phylogenetic analysis of a reduced sequence dataset. The resulting robust subfamilies were used in turn to generate Hidden Markov Models (HMMs), which will form the basis for the automated incorporation of new sequences into the continually expanding CAZy database.

RESULTS

Subfamily delineation

All-versus-all pairwise local sequence alignments were calculated for 22,946 GH16 domain sequences from the CAZy database in 210 minutes on a desktop computer (Intel Xeon Processor E5-1620 v4, 8 cores, 3.5 GHz, 16 GB RAM) For comparison, the computational time was reduced to 13 minutes using 128 cores on Compute Canada's WestGrid High Performance infrastructure. Subsequently, the BLAST result file was indexed over thresholds in intervals of 5 log units for *E*-values between 10^{-5} and 10^{-120}. Our preliminary SSN and HMM analysis indicated that the 10 SSNs for *E*-value thresholds between 10^{-20} and 10^{-65} were of most interest with the number of subfamilies ranging from 3 at $E = 10^{-20}$ to 27 at $E = 10^{-65}$ (Figure 2). Mapping sequence origin and the 15 currently known substrate specificities (from nearly 200 biochemically characterized GH16 proteins (7), Table 1) reveals the distribution of these features across emergent subfamilies (Figure 2).

To determine the threshold at which optimal discrimination of subfamilies is achieved, a library of HMMs was created for each SSN and their performance was evaluated by computing precision and recall rates using all 22,946 GH16 members as input (Figure 3). It was observed that at a threshold of $E = 10^{-60}$ the HMM library was able to retrieve all of the sequence assignments into the 26 subfamilies, with limited loss of precision at high *E*-values, compared to SSN based on lower thresholds (Figure 3). For SSNs induced by higher thresholds, GH16 was only broken-down into an increasing number of subfamilies, primarily along taxonomic lines (Additional file 1: Figure S1 and Figure S2). Such divisions are unlikely to be functionally significant and rather are likely only to reflect sequence drift due to speciation. In this analysis, it is also helpful to keep the limit-analysis in mind: division of GH16 into 22,946 individual subfamilies would result in recall and precision values of 100% at the subfamily level, yet it would provide no predictive power. Thus, although the data in Figure 3 would suggest that the HMM library from the SSN at $E = 10^{-60}$ may have the best performance, practically this represents little performance gain and might be unnecessarily stringent. Analysis of the taxonomic distribution and number of un-clustered sequences between the

SSN at $E = 10^{-60}$ and the previous SSN at $E = 10^{-55}$ suggest the latter would be a more pragmatic choice, considering that the continuous growth of GH16 family would likely result in new sequences filling the gaps between subfamilies that are too finely divided. Hence, the SSN at $E = 10^{-55}$ and the corresponding HMM library was chosen for the creation of the final subfamilies in GH16.

23 subfamilies were defined (Figure 4a), using the SSN based on the $E = 10^{-55}$ threshold of, which collectively assigned 22,367 sequences to a subfamily (97.5% of all GH16 modules analyzed). Family size ranges from 20 to 6,300 sequences. The taxonomical diversity within subfamilies mainly occurs at the phylum-level, with only four subfamilies (GH16_3, GH16_10, GH16_14, and GH16_21) present in multiple kingdoms of life. The lowest taxonomic diversity was in an early diverging group of mycobacterial sequences (GH16_9), which robustly formed a distinct subfamily (Figure 2). Notably, one of the earliest emerging features that distinguishes subfamilies is the presence or absence of the β-bulge sequence motif (EXDXXE vs. EXDXE) in the active-site β-strand presenting the catalytic residues (Figure 2), which is a key structural feature among GH16 members (30).

A limitation of SSNs is the inability to establish phylogenetic relationships between subfamilies. To establish overall context and to validate further the subfamily classification of GH16, a maximum-likelihood phylogenetic tree was constructed from 30 randomly selected sequences from each subfamily defined by the SSN. The delineation of subfamilies from the SSN (Figure 2 and Figure 4a), is identical to the monophyletic groups inferred from the phylogenetic tree (Figure 5a). Importantly, all clades comprising individual subfamilies are supported by high bootstrap values.

The SSN analysis delineated GH16 sequences into "characterized" subfamilies with one or more biochemically or structurally characterized members (denoted in the CAZy database (7)), and "uncharacterized" subfamilies for which structural-functional data is currently lacking. Table 1 summarizes the taxonomic range of source organisms, experimentally determined enzyme activities, and available tertiary structures for each subfamily shown in Figure 4a. Specific sequence accessions, including subfamily membership and characterization details, may be accessed directly in the CAZy database via URL http://www.cazy.org/GH16.html. In total, 16 of 23 subfamilies contained at least one biochemically characterized member, and 11 had a three-dimensional structure representative. Salient features of individual subfamilies are detailed below. Analogous to previous GH subfamily classifications (9, 10, 12), subfamilies are systematically referenced as "GH16_*n*", where *n* is the subfamily number.

Characterized subfamilies

GH16_1: The largest GH16 subfamily, GH16_1, has 6,300 members, which comprise almost exclusively fungal enzymes, with few members from a pathogenic nematode. GH16_1 is very distinct, already separating at a threshold of $E = 10^{-20}$ and exhibiting no significant segregation prior to a threshold of $E = 10^{-85}$ (Figure 4a and Additional file 1: Figure S1). Only fungal enzymes have been characterized in this subfamily: *endo*-β(1,3)-glucanases (EC 3.2.1.39) and *endo*-β(1,3)/β(1,4)-glucanases (EC 3.2.1.6) have been reported for 9 enzymes, while activity towards hyaluronate (hyaluronidase, EC 3.2.1.35) has been reported for 2 enzymes. Interestingly, one representative of this subfamily has been reported to be an *exo*-β(1,3)-glucosyltransferase/elongating β-transglucosylase (EC 2.4.1.-).

Structurally, GH16_1 is defined by the presence of numerous helical elements on the core β-jelly-roll fold, two in the N-terminal region and four in the C-terminal region, most of which are located on the opposite side of the structure from the active site cleft (Figure 5b). The α5 helix carries a conserved tryptophan, W257 (PDB ID: 2CL2), that points into the active site and faces a loop, which is consolidated by a disulfide bridge. Together, these elements define the positive enzyme subsites (31) in this subfamily. A notable sequence pattern "WPA....WPX" (X is often Y or N, but also A, T or I) is shared with GH16_3 and GH16_9 members. The "WPX" motif is located in a loop bordering the active-site cleft at the negative subsites and therefore likely contributes to substrate specificity.

GH16_2: Members of GH16_2 are almost exclusively reported in fungi, with less than 2% of the members found in plant-damaging oomycetes

(water molds) and algae. GH16_2 is very distinct and shows almost no sequence diversity even at a threshold of $E = 10^{-120}$ (Additional file 1: Figure S1 and Figure S2). Only a single biochemically characterized member, a cell-wall active β(1,6)-glucanase/transglucosylase (EC 3.2.1.-/2.4.1.-) from *Saccharomyces cerevisiae* is known (32). Interestingly, GH16_2 members lack a signal peptide that is otherwise commonly associated with members of fungal GH16 subfamilies. No tertiary structural representatives currently exist in GH16_2.

GH16_3: Historically known as the laminarinase subfamily (30, 33), GH16_3 is a large and extremely sequence-diverse subfamily (Figure 4a) found in all kingdoms. *Endo*-β(1,3)-glucanase (EC 3.2.1.39) and/or *endo*-β(1,3)/β(1,4)-glucanase activity (EC 3.2.1.6) has been reported in members of the Metazoa, Fungi, Archaea, and Bacteria. The broad taxonomical diversity of GH16_3 members makes this subfamily particularly sensitive to the threshold *E*-value cut-off, such that increasingly strict cut-off values result in fragmentation along taxonomic lines.

The large sequence and taxonomic diversity is reflected by low structural homology in this subfamily, where only very few stretches and features are strictly conserved among the subfamily members. However, the sequence pattern "WPA....WXX....WPX" (X being M or L for the second motif and A, K, R, M, or L after the third motif), similar to that found in GH16_1, is largely conserved throughout members of this subfamily. In GH16_3, this loop faces a short, subfamily-specific π-helical element that is located in the N-terminal region (residues 25 to 34 in PDB ID 4CTE). Furthermore, a tryptophan or phenylalanine, that lines the active-site in the positive subsites, is part of a partially conserved motif present in many subfamily members, as is a loop (H155 to H163) that contains a strongly conserved histidine residue (H155) facing this aromatic side chain. A structurally conserved short helical segment in different GH16_3 members (210 to 218) is located next to this loop and possibly participates in shaping the overall active-site cleft of GH16_3.

GH16_4: GH16_4 can be considered as a subfamily derived from GH16_3, which segregates along with GH16_5 and GH16_6 at lower *E*-value thresholds (Figure 2 and Figure 4b). GH16_4 contains members from the Metazoa and Fungal kingdoms, with *endo*-β(1,3)-glucanase (EC 3.2.1.39) activity reported for 13 enzymes from Metazoa. Significantly, about 9% of the 1900 GH16_4 members, across Metazoa and Fungi, have lost one or both of their catalytic residues, though this feature is not resolved into monophyletic groups in a phylogenetic analysis (data not shown). In comparison, this is the case for only 0.7% of GH16_3 members and 1% of all other GH16 members. No tertiary structural representatives currently exist in GH16_4.

GH16_8: One enzyme in the GH16_8 subfamily has been demonstrated to have *endo*-β(1,4)-galactosidase activity (EC 3.2.1.-) (34). The members of this subfamily have very high sequence similarity (no fragmentation in the SSN from $E = 10^{-40}$ to 10^{-120}, Figure 2, Additional file 1: Figure S1, and Figure S2), despite having members from both Firmicutes and Actinobacteria. About 75% of GH16_8 enzymes are linked to CBM32, members of which are known to bind galactose and are associated with wide variety of other GH domains. No tertiary structural representatives currently exist in GH16_8.

GH16_9: GH16_9 is comprised entirely of members from *Mycobacteria*. Although this observation contravenes our usual strict requirement for taxonomic diversity to establish a subfamily, the early segregation of this group at comparatively high *E*-values (Figure 2) supports the creation of a robust subfamily. Presently, no GH16_9 members have been biochemically characterized, but five members have been structurally characterized.

A structural characteristic of this subfamily is the low content of helical elements (Figure 5b), in which only a short helix is present in the N-terminal region adjacent to the first loop near the negative subsites. Remarkably, GH16_9 members generally lack aromatic residues in the negative subsites as compared to other subfamilies. A tryptophan (W154 in PDB ID 4PQ9) present in a conserved loop is positioned to accommodate a substrate in the positive subsites. Additionally, a conserved histidine (H161), which is also present in GH16_1, GH16_3, GH16_16, GH16_11, GH16_17, and

GH16_12, is found on the β-strand next to the catalytic EXDXXE motif.

GH16_10: Endo-β(1,3)-galactanases are exclusive to subfamily GH16_10, members of which have very high sequence similarity (SSN analysis indicates a stable group until a threshold cut-off of $E = 10^{-85}$, Figure 4a). Strikingly, this similarity is maintained across a wide taxonomic diversity, including the bacterial phyla Actinobacteria and Bacteroidetes, the fungal phyla Ascomycota and Basidiomycota, and the early diverging fungal lineage Chytridiomycota. *Endo*-β(1,3)-galactanase activity has been reported twice in Ascomycota species and once in Basidimycota, while the bacterial members remain to be biochemically characterized. No tertiary structural representatives currently exist in GH16_10.

GH16_11: GH16_11 is composed exclusively of bacterial members from the phylum Bacteroidetes, except for one member from *Coraliomargarita*, a bacterial member of the phylum Verrucomicrobia. The activity in GH16_11 is defined based on a single biochemically characterized β-porphyranase (EC 3.2.1.178).

Some key structural features of GH16_11 are shared with the β-agarase (GH16_15 and GH16_16), the β-porphyranase (GH16_12), and the κ-carrageenase (GH16_17) subfamilies, which is consistent with their close phylogenetic relationships (Figure 5a and Additional file 1: Figure S3). These subfamilies have a characteristic N-terminal feature that consists of a short β-strand followed by a helical element, which is not present in other GH16 members. GH16_11 is distinguished further by the spatial organization of the first loop bordering the negative subsites of the active-site cleft, as well as a conserved loop close to the C-terminus. This loop contains a characteristic arginine residue (R70 in PDB ID 3JUU) in addition to a conserved tryptophan W67 that is also present in GH16_16, both of which are involved in substrate binding. The loop formed by a conserved sequence motif close to the C-terminus (residues 256 to 265 in PDB ID 3JUU) is also structurally distinct from those in other subfamilies.

GH16_12: Like GH16_11, GH16_12 is composed exclusively of bacterial members from the Bacteroidetes phylum, except for one member from *Coraliomargarita* (Verrucomicrobia). GH16_12 contains three biochemically characterized β-porphyranases (EC 3.2.1.178). GH16_11 and GH16_12 are highly related and form a uniform subfamily at lower thresholds, precisely resolving into two subfamilies at the SSN threshold of $E = 10^{-55}$ (Figure 2 and Figure 4).

Consistent with the high relatedness of the two subfamilies, the major characteristic structural features are shared between the two subfamilies, including the N-terminus and the first loop bordering the negative subsites. GH16_12 is distinguished by specific amino acid substitutions in the aromatic platform of the −1 subsite, as well as various loops throughout the tertiary structure. Specifically, a loop between the C-terminal two β-strands, shared with GH16_11 is distinguished by sequence motives that are not identical between the two subfamilies, namely the stretch from 221 to 230 is "WNPVPKDGGM" in 3JUU, while the structural identical stretch from 288 to 297 is "WEKQVPTAED" in 4AWD. Additionally, the motif comprising residues 210 to 228 (PDB ID 4AWD), which in many other subfamilies forms a β-strand that terminates the β-sheet at the positive subsites, has a characteristic structure in GH16_12 members that begins at the level of the inner concave β-sheet at the positive subsites and then changes level to spatially board the outer β-sheet of the β-jelly-roll fold.

GH16_13: GH16_13 comprises sequences from marine bacteria and is the newest subfamily to have its activity revealed by biochemical characterization. One biochemical characterized member shown to hydrolyze furcellaran, a hybrid carrageenan containing both β-carrageenan and κ/β-carrageenan motifs (35). This subfamily has wide taxonomic distribution in the bacterial kingdom. No tertiary structural representatives currently exist in GH16_13.

GH16_15: Two β-agarases (EC 3.2.1.81) have been reported in the small GH16_15 (currently 24 members). This subfamily is very distinct from the other β-agarase-containing subfamily, GH16_16 (Figure 2), to which it forms a sister clade with high bootstrap support (Figure 5a). A member of GH16_15 has recently been shown to hydrolyze specifically complex agars from *Ceramiales* species, functionally distinguishing this subfamily

from GH16_16 (36). Notably, unlike GH16_16, no CBMs are associated with GH16_15.

Together with functional characterization, the first structural representative of GH16_15 has recently been solved (PDB ID 6HY3; (36)). This structure reveals high structural similarity with GH16_16, with differences mainly observed in specific amino acid substitutions. Particularly notable are two aromatic residues (W110 and Y112 in PDB ID 6HY3), which are located in the negative binding subsites, and a characteristic loop (residues 291−300) located near the positive binding subsites, which presents two tryptophan residues (W291 and W297) that point into the active-site cleft. Another unique feature of GH16_15 is the presence of a conserved arginine (R186) near the active site EXDXXE motif, as well as a second strictly conserved arginine (R224) located in the positive subsites.

GH16_16: Considering the size of GH16_16 (153 sequences), it is the most densely studied subfamily in GH16 with 32 biochemically characterized β-agarases (EC 3.2.1.81) from Bacteroidetes, Proteobacteria, and Actinobacteria. A CBM13 or CBM6 is found associated with approximately half of the GH16_16 members.

In GH16_16 a characteristic N-terminus is followed by an α-helix (G94−E99 in PDB ID 4ATF). This helix is not directly bordering the active site groove, however, it is immediately followed by a GH16_16-specific loop that contains a well-conserved tryptophan residue (W109) constituting subsite −3. Another characteristic feature of GH16_16 is the C-terminal motif from residues 308 to 315 that also presents an α-helix providing a tryptophan that forms the +3 subsite. Opposite of this feature is a loop including residues H215−F222, which contains a strictly conserved arginine residue (R219) that is involved in binding the 3,6-*anhydro* bridge of agarose in subsite −2.

GH16_17: GH16_17 contains κ-carrageenases (EC 3.2.1.83) from both Proteobacteria and Bacteroidetes. GH16_17 is the most distinct subfamily among those that hydrolyze marine carbohydrates, as it segregates at comparatively high *E*-value thresholds (Figure 2 and Figure 4). Examination of sequence subgroups in this subfamily highlights how sequence differences due to speciation can give the appearance of further subfamilies without a functional basis. The SSN sub-clusters (Figure 4) and phylogenetic clades (Figure 5) correspond roughly to taxonomic sub-divisions. Two members from different sub-branches have been structurally and biochemically analyzed, indicating that subtle differences in substrate recognition and mode of action (perhaps even life-style of the organism) are the result of evolutionary drift, while substrate specificity have remained constant – both are clearly kappa-carrageenases (37).

Despite the observed phylogenetic divergence from the β-agarases (GH16_15 and GH16_16) and the β-porphyranases (GH16_12), subfamily GH16_17 contains a similar, characteristic N-terminal spatial arrangement (Figure 5b). Otherwise, a key feature of this subfamily is vast diversity where only few elements are strictly conserved. A notable differentiator is found in the loop that follows the conserved tryptophan comprising the −1 subsite, which contains a well-conserved tyrosine or phenylalanine (Y143 in PDB ID 5OCR) that provides a hydrophobic environment to accommodate the 3,6-*anhydro* bridge in the −2 subsite. Importantly, a loop that is stabilized through two anti-parallel β-strands is positioned directly above the −1 subsite, thereby providing a strictly conserved arginine (R263) to bind the κ-carrageenan-specific sulfate group on O4 of galactose residues. GH16_17 have sequence variation around the positive subsites, indicating that subtle differences in substrate specificity might be found among this divergent subfamily.

GH16_18: GH16_18 is a large subfamily with 2,576 members. The subfamily is entirely composed of fungal enzymes including biochemically characterized chitin β(1,3)/β(1,6)-glucosyltransferases (EC 2.4.1.-) and cell-wall modifying enzymes (EC 3.2.1.-/2.4.1.-).

GH16_18 have a characteristic N-terminus, starting with a disulfide bridge (residues 25−40 in PDB ID 5NDL), which is arranged into a triple-stranded β-sheet with the C-terminus. Strikingly, no residues from this loop appear to participate to substrate binding in the negative subsites. On the other hand, one strictly conserved tryptophan, W207, forms a platform at the −2 subsite and the positive subsites also contain one strictly conserved tryptophan

residue (W221) and two largely conserved aromatic residues (F137 and Y145) that form large hydrophobic platforms to accommodate the substrate. W221 is situated in a subfamily-specific α-helix, α1, which is the only true α-helix present in GH16_18 members. Although F137 and Y145 are not strictly conserved, the loop that contains these residues is characteristic and largely conserved within GH16_18 members.

GH16_19: GH16_19 derives as a sister clade to GH16_18 (Figure 5a) and is composed of fungal enzymes, including a biochemically characterized chitin β(1,3)/β(1,6)- glucosyltransferase (EC 2.4.1.-). Notably, many fungi have orthologs in both GH16_18 and GH16_19. Apart from statistically significant sequence differences in the GH16 module, a major difference between the two subfamilies is the presence of a CBM18 (predicted to bind chitin) in practically all enzymes of GH16_19, whereas no CBM is associated with GH16_18. No tertiary structural representatives currently exist in GH16_19.

GH16_20: GH16_20 is a well characterized subfamily composed of plant enzymes specific for xyloglucan (38). Members of this subfamily are either xyloglucan *endo*-transglycosylases (XET, EC 2.4.1.207) or xyloglucan *endo*-hydrolases (XEH, EC 3.2.1.151) (28, 39).

A significant key feature of GH16_20 is the addition of a large C-terminal domain (residues 232−264 in PDB ID 2VH9; InterPro and PFAM "XET_C") that extends the active-site cleft at the positive subsites. In addition, a well-conserved loop region (residues 181−190) is located immediately adjacent to the catalytic residues and provides a strictly conserved tryptophan (W185) that forms a hydrophobic platform at the +1 subsite. At the negative subsites, the loops bordering the active-site cleft are characteristically short in GH16_20 members (40). The resulting broadening of the active-site cleft appears to be responsible for the recognition of the highly branched xyloglucan chain (41, 42). One exception is the loop that precedes the β-strand containing the catalytic EXDXE motif, which is specifically lengthened in the xyloglucan *endo*-hydrolases (28). Notably, the aromatic platform of the −1 subsite in GH16_20 members is a tyrosine (Y81), rather than a tryptophan found in most other GH16 members.

GH16_21: Historically known as the licheninase (EC 3.2.1.73) subfamily (30, 43), this subfamily has more than 30 biochemically characterized representatives among bacteria. Interestingly, a few members are found in the early diverging fungal lineage Chytridiomycota, including one biochemically characterized *endo*-β(1,3)/β(1,4)-glucanase (44). The *endo*-β(1,3)/β(1,4)-glucanases in GH16_21 strictly hydrolyze only the β(1,4)-glucosidic linkage in mixed-linkage β-glucan, typically at the anomeric position of backbone glucosyl units bearing a β(1,3) glucan kink, and do not hydrolyze β-glucans containing only β(1,3)- or β(1,4)-linkages. Thus, GH16_21 are functionally different from the *endo*-β(1,3)/β(1,4)-glucanases found in GH16_3, which hydrolyze β(1,3)- or β(1,4)-linkages in mixed-linkage β-glucan as well as β-glucans with only β(1,3)-linkages, such as laminarin.

Members of GH16_21 are among the shortest sequences, at about 210 residues, while the average length of most of the other GH16 proteins is 240 residues. Consequently, characteristic features of this subfamily are short loops surrounding the substrate binding groove. The conserved stretches are concentrated in four regions that border the central cleft, two on each side, which contain aromatic residues important for substrate binding (Y24, Y94, W103 and W192 in PDB ID 1GBG). Two of the characteristic loops contain short helical segments; the first (residues 91−100) is located at the −1 subsite, directly preceding the active site EXDXE motif, while the second borders the active-site on the opposite side (residues 189−193), thereby providing a strictly conserved tryptophan at the +1 subsite. In addition, and similar to the GH16_20 subfamily, the aromatic platform at the −1 subsite in GH16_21 members is a phenylalanine (F92), not a tryptophan.

Uncharacterized subfamilies

Six well-defined subfamilies currently await definition of biochemical activity (Table 1, Figure 2). In particular, two very large subfamilies of fungal origin, the two sister subfamilies GH16_22 and GH16_23, which collectively contain ca. 700 sequences, have so far gone unstudied. Likewise, two sister subfamilies, GH16_5 and GH16_7, limited to Proteobacteria, as well as GH16_6 with bacterial members, also remain unexplored.

Noteworthy is the early diverging subfamily GH16_12, a sister clade to the newly discovered GH16_13 furcellaranases that, despite few members, has high taxonomic diversity (Figure 2 and Figure 5a).

Non-classified sequences

Roughly 3% of the analyzed GH16 sequences were not assigned to subfamilies (Figure 2), primarily due to lack of a sufficient number of orthologs in the CAZy database to define a subfamily with at least 20 members and sufficient taxonomical diversity. Among these is the only characterized GH16 member from a virus (*Paramecium bursaria* Chlorella virus 1, GenBank AAC96462.1), which is an *endo*-β(1,3)/β(1,4)-glucanase that is distant from, but most closely related to, members of subfamily GH16_3. Other examples include two small groups related to the GH16_11 and GH16_12 β-porphyranase subfamilies, containing eight members and one biochemically characterized β-porphyranase each: β-porphyranase A (PDB ID 3ILF and 4ATE) (45, 46) and β-porphyranase C, respectively, from *Zobellia galactanivorans* DsijT. It is anticipated that these orphan sequences may seed additional subfamilies as the number of sequences in GenBank, from which the CAZy database is derived, continues to grow (7).

DISCUSSION

Advantages and limitations of SSN-based subfamily classification

The utilization of a Sequence Similarity Network-based approach allowed the division of 22,946 GH16 catalytic modules into subfamilies in a scalable, computationally efficient manner. Comparatively rapid generation of an all-versus-all pairwise scoring matrix, facile generation of SSNs at increasing BLAST *E*-value thresholds, and analysis of precision and recall rates, guided the selection of an SSN cut-off value producing 23 robust subfamilies (Figure 4a and Figure 2). A particular advantage of the SSN-based approach, versus classical phylogenetic methods based on MSAs, is the ability to utilize the full sequence dataset without the need for down-sampling to reduce computation time.

For example, the previous division of GH5 (9) and GH43 (12) into subfamilies based on molecular phylogeny, coped with the large amount of sequences (2,333 and 4,455, respectively) by employing the common practice of initial clustering of similar sequences, using algorithms such as UCLUST and CD-Hit (47, 48), to reduce the datasets. The clustering percent identity limitation for UCLUST and CD-hit are 50% and 40%, respectively, thus, in order to obtain a reliable clustering, percent identify cut-offs are usually set at 75% or higher (9, 12). In our preliminary analyses, applying a clustering cut-off of 75% to the 22,946 GH16 sequences yielded a reduced dataset of 7,557 sequences, which is still an order of magnitude larger than the dataset limitations for highly accurate MSA (49, 50) and subsequent maximum-likelihood phylogenetic tree estimation (51). Thus, a significant advantage of SSN generation is the superior computational efficiency due to fundamental differences in algorithm complexity compared to phylogenetic approaches. This allowed us to analyze the entire, unreduced GH16 dataset, which is 5-, 10-, and 13-times larger, respectively, than those used to classify GH43, GH5, and GH13 into subfamilies (9, 10, 12). Not least, a significant advantage of the combined BLAST-SSN approach is that it allows immediate recall of exact sequences from the dataset, including their precise location within the SSN, at any time, whereas individual sequence information is lost in phylogenies based on representative sequences.

On the other hand, SSNs are unable to establish unambiguous evolutionary relationships between subfamilies. As observed for the SSN at $E = 10^{-55}$ (Figure 4a), which we use to define GH16 subfamilies, there is no inter-subfamily connectivity, while at a relaxed threshold of $E = 10^{-25}$ the SSN reveals only the most basic relationships (Figure 4b). For example, GH16_17, which contains the marine carbohydrate-active κ-carrageenases, shows no connectivity to the other marine polysaccharidase subfamilies GH16_16, GH16_11, GH16_13, GH16_14, and GH16_15 at $E = 10^{-25}$, while these subfamilies appear to be connected to more evolutionarily distant subfamilies (30), *e.g.* GH16_3 (comprising terrestrial *endo*-β(1,3)-glucanases and *endo*-β(1,3)/β(1,4)-glucanases, Figure 4b). In contrast, a representative phylogenetic tree (Figure 5a) clearly indicates that the κ-carrageenases form a sister clade to the other marine subfamilies, in agreement

with a previously proposed evolution of GH16 diversity (30).

A roadmap for functional glycogenomics

The delineation of large families such as GH16 into subfamilies can greatly improve predictive power to guide future functional analyses of individual family members, as has been previously exemplified for GH5 (9), GH13 (10), GH43 (12), and the Polysaccharide Lyase families (13). In particular, subfamily association can provide strong suggestions of likely substrates, or substrate families, which should be prioritized in biochemical assays. Not least, subfamilies with no, or very few, functionally characterized members hold significant untapped potential for biochemical discovery. Together, the continued exploration of "known" and "unknown" subfamilies will continue to refine understanding of protein structure-function relationships across the evolutionary landscape of GH16.

In such endeavors, and especially for unsupervised bioinformatics, it is essential to bear in mind that this subfamily classification has certain predictive limitations. Sequence-alignment-based approaches to delineate subfamilies, including both SSN and phylogenetic approaches, have insufficient resolution to segregate sequences differing by minor variations, which may nonetheless have large effects on biochemical and biological function. For example, it is well known that single amino acid substitutions can switch substrate specificity in glycosidases (52, 53).

Within GH16 subfamilies, such limitations are exemplified by several cases. Neither SSNs (Figure 4a) nor phylogeny (Figure 5a) allow for the segregation the β(1,3)-glucanases in GH16_4 from the homologous non-catalytic binding proteins, in which the catalytic residues are mutated, even at very high threshold values ($E > 10^{-85}$, Additional file 1: Figure S1 and Figure S2). GH16_3 is known to comprise both *endo*-β(1,3)-glucanases (laminarinases, EC 3.2.1.39) and *endo*-β(1,3)/β(1,4)-glucanases (the latter hydrolyzing the β(1,4)-bond in mixed-linkage glucan, EC 3.2.1.73) (54), which likewise do not segregate cleanly in SSNs nor phylogenies. Lastly, the canonical double-displacement mechanism of GH16 enzymes allows for both glycosyl transfer to water (hydrolysis, EC 3.2.1.-) and/or carbohydrate acceptor substrates (transglycosylation, EC 2.4.1.-) (55). The subfamily classification described here does not segregate transglycosylases from hydrolases in four fungal subfamilies (GH16_1, GH16_2, GH16_18, and GH16_19) and one plant subfamily (GH16_20), (28), Table 1, indicating that the determinants of such specificities represent weak sequence signals masked by background sequence noise.

In light of current rapid increases in sequence data volume and a comparatively limited amount of experimental CAZyme characterization, there is significant potential for the propagation of inaccurate functional annotations due to overconfident bioinformatic assignments. Consequently, this jeopardizes the usefulness of such annotations. We therefore advocate a conservative approach, in which functional predictions are abandoned altogether in (meta)genomic sequence annotation, in favor of simply designating all predicted proteins by their family and subfamily numbers, e.g. GH16_*n*.

The evolution of structure-function relationships in GH16

At the highest level, this subfamily classification enables the evolution of major structural features to be mapped across GH16. Generally, variability within a subfamily is concentrated in the loops connecting the β-strands of the concave β-sheet (forming the active site groove), rather than in the N-terminal or C-terminal regions. In contrast, the termini typically vary substantially between subfamilies (Figure 5b), e.g. the additional N- and C-terminal helices in GH16_1 or the expanded C-terminus in GH16_20, which have significant functional ramifications (28).

Interestingly, some large subfamilies are highly conserved, such as the mycobacterial-specific GH16_9 subfamily and the plant-specific GH16_20 XTHs, whereas some smaller subfamilies, such as the GH16_16 β-agarases and GH16_17 κ-carrageenases, display substantial variability, even though they appear to display the same global substrate specificity (within the limits of current biochemical characterization). This might be related to specific constraints with respect to their biological functions. For example the crucial biological role of GH16_20 xyloglucan endo-transglycosylases and endo-hydrolases in plant

growth and development (22, 38) might constrain sequence variations, whereas the bacterial catabolic enzymes may have diversified as a consequence of adaptation to available substrate diversity and environmental niches (2, 3, 56, 57). If this hypothesis holds true for the currently uncharacterized mycobacterial GH16_9 enzymes, a crucial biological role of the GH16 enzymes for these organisms can be expected.

Looking to the future: Emerging subfamilies

The CAZy database is derived exclusively from the NCBI Genbank daily releases for practical reasons (7). Consequently, CAZy database, and thus the entire GH16 sequence set used here, does not capture sequences from nascent (meta)genomic efforts, especially unfinished genomes from sequencing center databases (*e.g.*, Joint Genome Institute, Broad Institute, Beijing Genomics Institute, *etc.*). Thus, it can be reasonably anticipated that the number of GH16 subfamilies will increase beyond the 23 presented here as the number of sequences in Genbank continues to increase. This includes subfamilies from currently identified groups with fewer than 20 sequences or currently low taxonomic diversity, as well as newly emergent subfamilies from currently unexplored sequence space.

An example of an emerging GH16 subfamily is comprised of recently identified mixed-function *endo*-β(1,3)/β(1,4)-glucanases/*endo*-xyloglucanases from plants, for which biochemical and structural information exists (*e.g.*, PDB ID 5DZF and 5DZG). These EG16 (endo-glucanase, GH16) members represent functional intermediates and an evolutionary link between the classic bacterial *endo*-β(1,3)/β(1,4)-glucanases in GH16_21 and the plant xyloglucan endo-transglycosylases and endo-hydrolases in GH16_20 (41, 42). A comprehensive census using genomes and transcriptomes of over 1,200 plant species has revealed a large collection of EG16 sequences in plant sequence databases, which are currently not deposited in Genbank (23). Generation of SSNs including 717 plant EG16 orthologs with the 22,946 CAZy GH16 entries indicated that EG16 members segregate from GH16_20 at a threshold between $E = 10^{-35}$ and $E = 10^{-40}$ (data not shown), and thus will form an independent subfamily in the future. This subfamily was verified by Maximum Likelihood phylogeny, in which EG16 members constitute a sister group to the xyloglucan endo-transglycosylases and endo-hydrolases with high bootstrap support (Figure 5a).

CODA

Since the introduction of protein SSN analysis in its present form a decade ago (16), the use of SSNs has been growing in popularity for the analysis of large datasets (17, 58–66), due in part to a lower computational demand than classical molecular phylogeny. Here, we have utilized the power of SSN analysis to devise a robust subfamily classification of the large and diverse family GH16. This framework, which collates biochemical and structural data on characterized members, will enable more refined functional prediction to guide future bioinformatics and experimental studies. Nonetheless, we advocate a conservative approach to protein annotation, in which uncharacterized enzymes are referred to solely by their subfamily membership, to avoid the propagation of misleading functional annotation in public databases. To aid future sequence annotation, the GH16 subfamily classification is now publicly available in the CAZy database via URL http://www.cazy.org/GH16.html.

EXPERIMENTAL PROCEDURES

Data acquisition

All GH16 members were extracted from the CAZy database (February 2018) (7), and used to retrieve amino-acid sequences from GenBank. During this step, additional meta-information was gathered, including taxonomic lineage (Kingdom, Phylum, Class, Order, Family, Genus and Species ranks), modularity (presence of CBMs, signal peptides, etc.) of the full length sequence (semi-manually annotated using in-house CAZy pipelines (67)), and both biochemical and structural information from the literature. Sequences with less than 95% coverage to the GH16 family model were considered as fragments (13.7% in total) and not included in the final dataset.

Sequence Similarity Network analysis

All-versus-all pairwise local alignments of the 22,946 GH16 domain protein sequences were computed by BLAST+ 2.2.31 with default settings (specifically, Scoring Matrix: BLOSUM62, Gap opening: 11; Gap extension: 1) (68) using GNU

Parallel (69), which generated the *E*-value, bit score, alignment length, sequence identity, and sequence similarity for sequence pairs. The data were filtered using specific *E*-value threshold cut-offs (from least stringent, $E = 10^{-5}$, to most stringent, $E = 10^{-120}$) to generate a series of associated SSNs. To formally constitute a subfamily, connected clusters were required to contain at least 20 sequences, which ensured diversity above the taxonomic *class* level to mitigate against bias arising from over-representation of closely related organisms and GH16 homologs (9–13). Members of each putative subfamily were identified using NetworkX (70). SSNs were visualized with Cytoscape (71) using the *yFiles organic* layout. To simplify the display of large SSNs, nodes representing highly similar sequences (*E*-value of $E = 10^{-85}$) were merged into meta-nodes using the *depth-first search* algorithm (72). The bioinformatics workflow used here has been packaged into a graphical user interface-based program, SSNpipe, which is freely available on GitHub at URL https://github.com/ahvdk/SSNpipe.

Subfamily assessment/validation using Hidden Markov Models

Each SSN, defined by its clustering threshold (BLASTP *E*-value), can be considered as a set of *N* assignments ($p \rightarrow s$), where *p*, each of the 22,946 proteins, *p*, is assigned to its subfamily, *s*, among *S* total subfamilies. HMMs for each subfamily in each SSN were used to measure precision and recall rates to assess SSN utility and validate the choice of an optimal threshold value for GH16, as follows:

A library of *S*+1 HMMs was assembled, corresponding to one HMM for each subfamily *s* and an additional HMM for the remaining GH16 members. Each HMM was generated using the *hmmbuild* command in HMMER3.2 with default parameters (73). Sequence sets were first reduced in redundancy (75%) using UCLUST (47), the resulting sequences were aligned with MAFFT using the G-INS-i strategy (iterative refinement, using WSP and consistency scores, of pairwise Needleman-Wunsch global alignments) (74), and each alignment was inspected in Jalview (75) to manually define the boundaries of the GH16 module.

The *hmmscan* command in HMMER3.2 was then used to search the 22,946 GH16 modules against the collection of *S*+1 HHMs. A protein *p'* was considered to belong definitively to a subfamily HMM, *s'*, only if (i) the best-matching HMM *E*-value was below 10^{-30} and (ii) the second best-matching HMM had an *E*-value at least 10^{-10} fold greater (i.e., less significant). The resulting set of *P* predictions ($p' \rightarrow s'$) was compared to the *N* reference assignments ($p \rightarrow s$) from the SSN. Identities between predictions and assignments were counted as true positives (*TP*). Predictions ($p' \rightarrow s'$) for a protein *p'* not initially assigned to the same subfamily or to any subfamily (GH16 members unclassified in a subfamily by the SSN) were counted as false positives (*FP*). The assignments ($p \rightarrow s$) for a protein *p* not predicted in any subfamily (GH16 unclassified at the subfamily level by the HMMs) are counted as false negatives (*FN*). To generate precision/recall plots, *precision* = $TP/(TP+FP)$ and *recall* = $TP/(TP+FN)$.

Molecular phylogeny

For each subfamily, 30 random sequences (or all sequences in subfamilies with less than 30 members) were aligned with MAFFT using the G-INS-i (Iterative refinement, using WSP and consistency scores, of pairwise Needleman-Wunsch global alignments) strategy (74). Three GH7 sequences (GenBank accessions CAA37878.1, ABY56790.1 and AAM54070.1) were included as an out-group. The quality of the alignment was ensured by manual inspection in Jalview (75) and corrected according to available structural information if necessary. A maximum-likelihood phylogenetic tree was estimated with RAxML (76) (100 bootstrap replicates) and visualized with iTOL (77).

Structural comparison

The crystal structure coordinates for forty-two GH16 members were downloaded from the Protein Data Bank (PDB), and pairwise superimposed starting from one of the shortest sequences (PDB ID: 1GBG) using the SSM algorithm (78) in Coot (79). One representative member was selected for those subfamilies where multiple structures are available (Table 1). For each subfamily, at least 10 randomly chosen sequences, in addition to that of the structural representative, were aligned with Multalin (80) and visualized adding the secondary structure elements using Espript (81). For each subfamily the superimposed coordinates were

visually inspected for conserved and divergent residues around the active site groove, the central −1 and +1 binding subsites, and conserved and characteristic features were highlighted in structural icons using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.).

ACKNOWLEDGEMENTS

Dima Vavilov (MSL, UBC) is acknowledged for technical assistance with accessing computational infrastructure. Work in Vancouver was supported by funding for the project "SYNBIOMICS - Functional genomics and techno-economic models for advanced biopolymer synthesis" from Genome Canada, with additional support from Ontario Genomics, Genome Quebec, and Genome BC (project #10405, www.synbiomics.ca). This research was supported by computational resources provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca). Work in Marseille was supported by grants ANR-14-CE06-0017 and ANR-17-CE20-0032 of Agence Nationale de la Recherche, France and by the Novozymes Prize awarded to BH by the Novo-Nordisk Foundation, Denmark. MC and GM acknowledge support from ANR via the investment expenditure program IDEALG (http://www.idealg.ueb.eu, grant agreement No. ANR-10-BTBR-04). Work in Roscoff was also funded by the European Union Horizon 2020 programme (project ID 727892, GenialG - GENetic diversity exploitation for Innovative Macro-ALGal biorefinery).

CONFLICT OF INTEREST

The Authors declare that there are no competing interests associated with the manuscript.

AUTHOR CONTRIBUTIONS

GM, HB, and MC conceived the study. VL, NT, and BH prepared the input protein sequence datasets. AHV and NT performed the computational analysis and subfamily delineation, with guidance by BH and HB. MC analyzed protein structures. AHV wrote the manuscript with input from HB, MC, NT, and BH. All authors read and approved the final manuscript.

REFERENCES

1. Varki, A. (2017) *Essentials of Glycobiology.*, 3rd Ed., Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press

2. Popper, Z. A., Michel, G., Hervé, C., Domozych, D. S., Willats, W. G. T., Tuohy, M. G., Kloareg, B., and Stengel, D. B. (2011) Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. *Annu. Rev. Plant Biol.* **62**, 567–590

3. Burton, R. A., Gidley, M. J., and Fincher, G. B. (2010) Heterogeneity in the chemistry, structure and function of plant cell walls. *Nat. Chem. Biol.* **6**, 724–732

4. Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*. **281**, 237–240

5. Bar-On, Y. M., Phillips, R., and Milo, R. (2018) The biomass distribution on Earth. *Proc. Natl. Acad. Sci.* **115**, 6506–6511

6. Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., Frederick Jr., W. J., Hallett, J. P., Leak, D. J., Liotta, C. L., Mielenz, J. R., Murphy, R., Templer, R., and Tschaplinski, T. (2006) The Path Forward for Biofuels and Biomaterials. *Science*. **311**, 484–489

7. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495

8. CAZypedia Consortium (2018) Ten years of CAZypedia: A living encyclopedia of carbohydrate-active enzymes. *Glycobiology*. **28**, 3–8

9. Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, H., and Henrissat, B. (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* **12**,

10. Stam, M. R., Danchin, E. G. J., Rancurel, C., Coutinho, P. M., and Henrissat, B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: Towards improved functional annotations of α-amylase-related proteins. *Protein Eng. Des. Sel.* **19**, 555–562

11. St John, F. J., González, J. M., and Pozharski, E. (2010) Consolidation of glycosyl hydrolase family 30: A dual domain 4/7 hydrolase family consisting of two structurally distinct groups. *FEBS Lett.* **584**, 4435–4441

12. Mewis, K., Lenfant, N., Lombard, V., and Henrissat, B. (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: A motivation for detailed enzyme characterization. *Appl. Environ. Microbiol.* **82**, 1686–1692

13. Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., and Henrissat, B. (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J.* **432**, 437–444

14. Liu, K., Linder, C. R., and Warnow, T. (2011) Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Curr.* **2**, RRN1198

15. Carrillo, H., and Lipman, D. (1988) The Multiple Sequence Alignment Problem in Biology. *SIAM J. Appl. Math.* **48**, 1073–1082

16. Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*. **4**, e4345

17. Copp, J. N., Akiva, E., Babbitt, P. C., and Tokuriki, N. (2018) Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry*. **57**, 4651–4662

18. El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D., and Henrissat, B. (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.* **11**, 497–504

19. Brouwer, H., Coutinho, P. M., Henrissat, B., and de Vries, R. P. (2014) Carbohydrate-related enzymes of important *Phytophthora* plant pathogens. *Fungal Genet. Biol.* **72**, 192–200

20. Zhao, Z., Liu, H., Wang, C., and Xu, J.-R. (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics*. **14**, 274

21. Cabib, E., Farkas, V., Kosík, O., Blanco, N., Arroyo, J., and McPhie, P. (2008) Assembly of the yeast cell wall: *Crh1p and Crh2p act as transglycosylases in vivo and in vitro*. *J. Biol. Chem.* **283**, 29859–29872

22. Rose, J. K. C., Braam, J., Fry, S. C., and Nishitani, K. (2002) The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: Current perspectives and a new unifying nomenclature. *Plant Cell Physiol.* **43**, 1421–1435

23. Behar, H., Graham, S. W., and Brumer, H. (2018) Comprehensive cross-genome survey and phylogeny of glycoside hydrolase family 16 members reveals the evolutionary origin of EG16 and XTH proteins in plant lineages. *Plant J.* **95**, 1114–1128

24. Hughes, A. L. (2012) Evolution of the βGRP/GNBP/β-1,3-glucanase family of insects. *Immunogenetics*. **64**, 549–58

25. Elyakova, L. A., and Shilova, T. G. (1979) Characterization of the type of action of β-1,3-glucanases from marine invertebrates. *Comp. Biochem. Physiol. Part B Comp. Biochem.* **64**, 245–248

26. Keitel, T., Simon, O., Borriss, R., and Heinemann, U. (1993) Molecular and active-site structure of a *Bacillus* 1,3-1,4-β-glucanase. *Proc. Natl. Acad. Sci.* **90**, 5287–5291

27. Hehemann, J.-H., Boraston, A. B., and Czjzek, M. (2014) A sweet new wave: structures and mechanisms of enzymes that digest polysaccharides from marine algae. *Curr. Opin. Struct. Biol.* **28**, 77–86

28. Baumann, M. J., Eklöf, J. M., Michel, G., Kallas, Å. M., Teeri, T. T., Czjzek, M., and Brumer, H. (2007) Structural Evidence for the Evolution of Xyloglucanase Activity from Xyloglucan *Endo* -Transglycosylases: Biological Implications for Cell Wall Metabolism. *Plant Cell*. **19**, 1947–1963

29. Lee, H., Kwon, H., Park, J., Kurokawa, K., and Lee, B. L. (2009) N-terminal GNBP homology domain of Gram-negative binding protein 3 functions as a β-1,3-glucan binding motif in *Tenebrio molitor*. *BMB Rep.* **42**, 506–510

30. Michel, G., Chantalat, L., Duee, E., Barbeyron, T., Henrissat, B., Kloareg, B., and Dideberg, O. (2001) The κ-carrageenase of *P. carrageenovora* Features a Tunnel-Shaped Active Site. *Structure*. **9**, 513–525

31. Davies, G. J., Wilson, K. S., and Henrissat, B. (1997) Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J.* **321**, 557–559

32. Bowen, S., and Wheals, A. E. (2004) Incorporation of Sed1p into the cell wall of *Saccharomyces cerevisiae* involves *KRE6*. *FEMS Yeast Res.* **4**, 731–735

33. Barbeyron, T., Gerard, A., Potin, P., Henrissat, B., and Kloareg, B. (1998) The Kappa-Carrageenase of the Marine Bacterium *Cytophaga drobachiensis*. Structural and Phylogenetic Relationships Within Family-16 Glycoside Hydrolases. *Mol. Biol. Evol*. **15**, 528–537

34. Ashida, H., Maskos, K., Li, S.-C., and Li, Y. (2002) Characterization of a Novel Endo-β-galactosidase Specific for Releasing the Disaccharide GlcNAcα1→4Gal from Glycoconjugates. *Biochemistry*. **41**, 2388–2395

35. Schultz-Johansen, M., Bech, P. K., Hennessy, R. C., Glaring, M. A., Barbeyron, T., Czjzek, M., and Stougaard, P. (2018) A Novel Enzyme Portfolio for Red Algal Polysaccharide Degradation in the Marine Bacterium *Paraglaciecola hydrolytica* $S66^T$ Encoded in a Sizeable Polysaccharide Utilization Locus. *Front. Microbiol.* **9**, 1–15

36. Naretto, A., Fanuel, M., Ropartz, D., Rogniaux, H., Larocque, R., Czjzek, M., Tellier, C., and Michel, G. (2019) The agar-specific hydrolase *Zg*AgaC from the marine bacterium *Zobellia galactanivorans* defines a new GH16 protein subfamily. *J. Biol. Chem.* **294**, 6923–6939

37. Matard-Mann, M., Bernard, T., Leroux, C., Barbeyron, T., Larocque, R., Préchoux, A., Jeudy, A., Jam, M., Nyvall Collén, P., Michel, G., and Czjzek, M. (2017) Structural insights into marine carbohydrate degradation by family GH16 κ-carrageenases. *J. Biol. Chem.* **292**, 19919–19934

38. Eklöf, J. M., and Brumer, H. (2010) The XTH Gene Family: An Update on Enzyme Structure, Function, and Phylogeny in Xyloglucan Remodeling. *Plant Physiol.* **153**, 456–466

39. Kaewthai, N., Gendre, D., Eklöf, J. M., Ibatullin, F. M., Ezcurra, I., Bhalerao, R. P., and Brumer, H. (2013) Group III-A XTH Genes of Arabidopsis Encode Predominant Xyloglucan Endohydrolases That Are Dispensable for Normal Growth. *Plant Physiol.* **161**, 440–454

40. Johansson, P., Brumer, H., Baumann, M. J., Kallas, A. M., Henriksson, H., Denman, S. E., Teeri, T. T., and Jones, T. A. (2004) Crystal structures of a poplar xyloglucan endotransglycosylase reveal details of transglycosylation acceptor binding. *Plant Cell*. **16**, 874–86

41. Eklöf, J. M., Shojania, S., Okon, M., McIntosh, L. P., and Brumer, H. (2013) Structure-function analysis of a broad specificity *Populus trichocarpa* endo-β-glucanase reveals an evolutionary link between bacterial licheninases and plant *XTH* gene products. *J. Biol. Chem.* **288**, 15786–15799

42. McGregor, N., Yin, V., Tung, C.-C., Van Petegem, F., and Brumer, H. (2017) Crystallographic insight into the evolutionary origins of xyloglucan endotransglycosylases and endohydrolases. *Plant J.* **89**, 651–670

43. Planas, A. (2000) Bacterial 1,3-1,4-β-glucanases: structure, function and protein engineering. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **1543**, 361–382

44. Chen, H., Li, X. L., and Ljungdahl, L. G. (1997) Sequencing of a 1,3-1,4-β-D-glucanase (lichenase) from the anaerobic fungus *Orpinomyces* strain PC-2: properties of the enzyme expressed in *Escherichia coli* and evidence that the gene has a bacterial origin. *J. Bacteriol.* **179**, 6028–6034

45. Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*. **464**, 908–912

46. Hehemann, J.-H., Correc, G., Thomas, F., Bernard, T., Barbeyron, T., Jam, M., Helbert, W., Michel, G., and Czjzek, M. (2012) Biochemical and Structural Characterization of the Complex Agarolytic Enzyme System from the Marine Bacterium *Zobellia galactanivorans*. *J. Biol. Chem.* **287**, 30571–30584

47. Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. **26**, 2460–2461

48. Li, W., and Godzik, A. (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**, 1658–1659

49. Le, Q., Sievers, F., and Higgins, D. G. (2017) Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*. **33**, 1331–1337

50. Yamada, K. D., Tomii, K., and Katoh, K. (2016) Application of the MAFFT sequence alignment program to large data - Reexamination of the usefulness of chained guide trees. *Bioinformatics*. **32**, 3246–3251

51. Liu, K., Linder, C. R., and Warnow, T. (2011) RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One*. **6**, e27731

52. Arnal, G., Stogios, P. J., Asohan, J., Skarina, T., Savchenko, A., and Brumer, H. (2018) Structural enzymology reveals the molecular basis of substrate regiospecificity and processivity of an exemplar bacterial glycoside hydrolase family 74 *endo*-xyloglucanase. *Biochem. J.* **475**, 3963–3978

53. Ichinose, H., Fujimoto, Z., Honda, M., Harazono, K., Nishimoto, Y., Uzura, A., and Kaneko, S. (2009) A β-L-Arabinopyranosidase from *Streptomyces avermitilis* Is a Novel Member of Glycoside Hydrolase Family 27. *J. Biol. Chem.* **284**, 25097–25106

54. Tamura, K., Hemsworth, G. R., Déjean, G., Rogers, T. E., Pudlo, N. A., Urs, K., Jain, N., Davies, G. J., Martens, E. C., and Brumer, H. (2017) Molecular Mechanism by which Prominent Human Gut Bacteroidetes Utilize Mixed-Linkage Beta-Glucans, Major Health-Promoting Cereal Polysaccharides. *Cell Rep.* **21**, 417–430

55. Sinnott, M. L. (1990) Catalytic mechanism of enzymic glycosyl transfer. *Chem. Rev.* **90**, 1171–1202

56. Xu, S. Y., Huang, X., and Cheong, K. L. (2017) Recent advances in marine algae polysaccharides:

Isolation, structure, and activities. *Mar. Drugs*. **15**, 1–16

57. Gow, N. A. R., Latge, J.-P., and Munro, C. A. (2017) The Fungal Cell Wall: Structure, Biosynthesis, and Function. *Microbiol. Spectr.* **5**, 188–192

58. Glasner, M. E. (2017) Finding enzymes in the gut metagenome. *Science*. **355**, 577–578

59. Levin, B. J., Huang, Y. Y., Peck, S. C., Wei, Y., Martínez-del Campo, A., Marks, J. A., Franzosa, E. A., Huttenhower, C., and Balskus, E. P. (2017) A prominent glycyl radical enzyme in human gut microbiomes metabolizes *trans*-4-hydroxy-L-proline. *Science*. **355**, eaai8386

60. An, L., Cogan, D. P., Navo, C. D., Jiménez-Osés, G., Nair, S. K., and van der Donk, W. A. (2018) Substrate-assisted enzymatic formation of lysinoalanine in duramycin. *Nat. Chem. Biol.* **14**, 928–933

61. Welsh, M. A., Taguchi, A., Schaefer, K., Van Tyne, D., Lebreton, F., Gilmore, M. S., Kahne, D., and Walker, S. (2017) Identification of a Functionally Unique Family of Penicillin-Binding Proteins. *J. Am. Chem. Soc.* **139**, 17727–17730

62. Jeoung, J.-H., and Dobbek, H. (2018) ATP-dependent substrate reduction at an $[Fe_8S_9]$ double-cubane cluster. *Proc. Natl. Acad. Sci.* **115**, 2994–2999

63. González, J. M., Hernández, L., Manzano, I., and Pedrós-Alió, C. (2019) Functional annotation of orthologs in metagenomes: a case study of genes for the transformation of oceanic dimethylsulfoniopropionate. *ISME J.* **13**, 1183–1197

64. Colin, P.-Y., Kintses, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., Hyvönen, M., Morgavi, D. P., Janssen, D. B., and Hollfelder, F. (2015) Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.* **6**, 10008

65. Benjdia, A., Guillot, A., Ruffié, P., Leprince, J., and Berteau, O. (2017) Post-translational modification of ribosomally synthesized peptides by a radical SAM epimerase in *Bacillus subtilis*. *Nat. Chem.* **9**, 698–707

66. Giessen, T. W., and Silver, P. A. (2017) Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nat. Microbiol.* **2**, 17029

67. Coutinho, P. M., Rancurel, C., Stam, M., Bernard, T., Couto, F. M., Danchin, E. G. J., and Henrissat, B. (2009) Carbohydrate-Active Enzymes Database: Principles and Classification of Glycosyltransferases. in *Bioinformatics for Glycobiology and Glycomics*, pp. 89–118, John Wiley & Sons, Ltd, Chichester, UK

68. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics*. **10**, 1–9

69. Tange, O. (2011) GNU Parallel: the command-line power tool. ;*login USENIX Mag.* **36**, 42–47

70. Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008) Exploring network structure, dynamics, and function using NetworkX. *Proc. 7th Python Sci. Conf.* **SciPy 2008**, 11–15

71. Shannon, P., Markiel, A., Owen Ozier, 2, Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504

72. Tarjan, R. (1972) Depth-First Search and Linear Graph Algorithms. *SIAM J. Comput.* **1**, 146–160

73. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013) Challenges in homology

search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121

74. Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780

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

76. Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313

77. Letunic, I., and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245

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

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

80. Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890

81. Robert, X., and Gouet, P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, 320–324

ABBREVIATIONS

Sequence Similarity Network, SSN; Hidden Markov Model, HMM; Maximum Likelihood, ML; Multiple Sequence Alignment (MSA); carbohydrate-active enzymes, CAZymes; glycosyltransferases, GT; glycoside hydrolase, GH; Glycoside Hydrolase Family 16, GH16; polysaccharide lyase, PL; carbohydrate esterase, CE; auxiliary activity enzymes, AA; carbohydrate-binding modules, CBM;

TABLE

Table 1. Defined subfamilies within GH16

#	Name	Taxonomical distribution	EC		Sequences (#)	Characterized members (#)	Representative PDB structure
1	FUN1	Eukaryota	3.2.1.39 3.2.1.6 3.2.1.35 2.4.1.-	*Endo*-β(1,3)-glucanase *Endo*-β(1,3)/β(1,4)-glucanase Hyaluronidase Transglycosylase	6300	13	2CL2
2	FUN2	Eukaryota	2.4.1.-/3.2.1.-	Transglycosylase	3422	1	
3	LAM1	Diverse	3.2.1.39 3.2.1.6	*Endo*-β(1,3)-glucanase *Endo*-β(1,3)/β(1,4)-glucanase	3749	38	4CTE
4	LAM2	Eukaryota	3.2.1.39	*Endo*-β(1,3)-glucanase	1896	13	
5	UNK3	Proteobacteria			115	0	
6	UNK4	Bacteria			31	0	
7	UNK5	Proteobacteria			51	0	
8	EGA	Bacteria	3.2.1.-	*Endo*-β(1,4)-galactosidase	41	1	
9	MB	*Mycobacterium*			346	0	4PQ9
10	GAL	Diverse	3.2.1.181	*Endo*-β(1,3)-galactanase	343	3	
11	POR1	Bacteria	3.2.1.178	β-porphyranase	52	1	3JUU
12	POR2	Bacteria	3.2.1.178	β-porphyranase	20	3	4AWD
13	FUR1	Bacteria	-	Furcellaranase	44	1	
14	UNK6	Diverse			28	0	
15	AGA2	Bacteria	3.2.1.81	β-agarase	24	2	6HY3
16	AGA1	Bacteria	3.2.1.81	β-agarase	153	32	4ATF
17	CAR	Bacteria	3.2.1.83	κ-carrageenase	38	6	5OCR
18	CHI1	Fungi	2.4.1.- 2.4.1.-/3.2.1.-	Chitin β(1,6)-glucanosyltransferase Cell-wall modifying	2576	2	5NDL
19	CHI2	Fungi	2.4.1.-	Chitin β(1,6)-glucanosyltransferase	1129	1	
20	XTH	Plantae	2.4.1.207 3.2.1.151	Xyloglucan *endo*-tranglycosylase Xyloglucan *endo*-hydrolase	719	34	2VH9
21	LIC	Diverse	3.2.1.73	*Endo*-β(1,3)/β(1,4)-glucanase	647	35	1GBG
22	UNK1	Fungi			555	0	
23	UNK2	Ascomycota			119	0	

FIGURES AND FIGURE LEGENDS

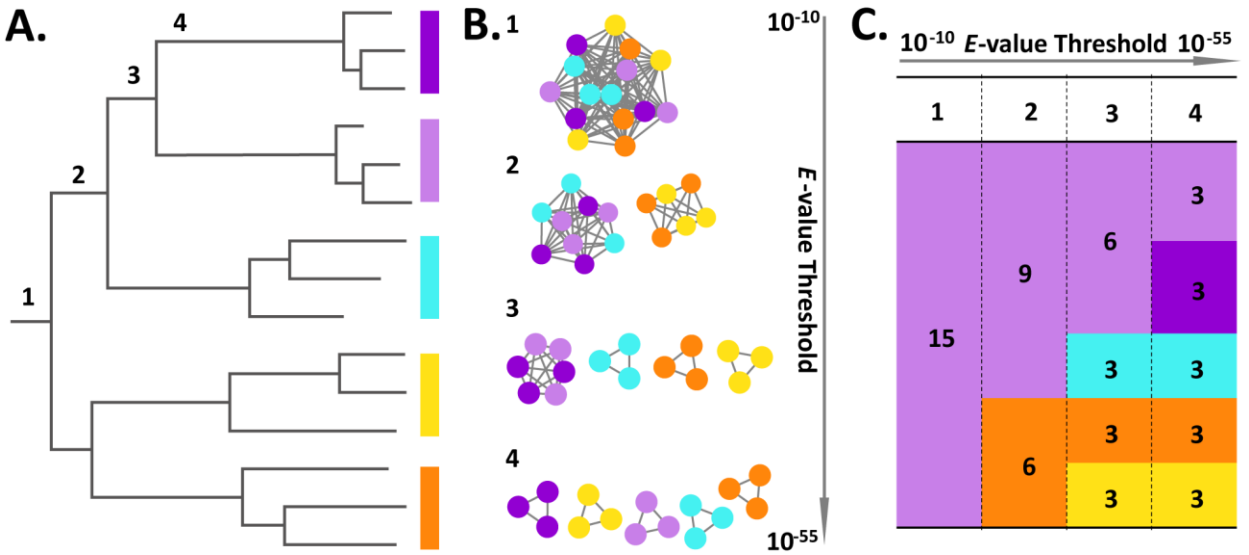

Figure 1. Subfamily delineation based on distinct analysis/representation. This artificial example of 15 sequences to be classified into subfamilies, illustrates the relationships between distinct representation and analysis. The numbers 1−4 indicate four hypothetical subfamily classifications which are concordant in all three representation. **a. Evolutionary tree:** reconstruction from a phylogenetic analysis or hierarchical clustering. Subfamily delineation consists in drawing a vertical line (below 1-4 numbers) and make a family for each out-coming branch **b. SSN connection graph:** Sequence Similarity Networks (SSNs) with sequences represented as nodes (circles) and all pairwise sequence relationships (alignments) above a defined *E*-value threshold indicated with edges (lines). At increased thresholds (1-4 numbers), the connected components break up into an increasing number of subcomponents, representing putative subfamily delineations. **c. SNN tabular summary:** each column (1-4 number for each *E*-value threshold, separated by a vertical dashed line) depicts a distinct subfamilization and displays the number of clusters/subfamilies as colored boxes, and the number of members/sequences in each cluster/subfamily.

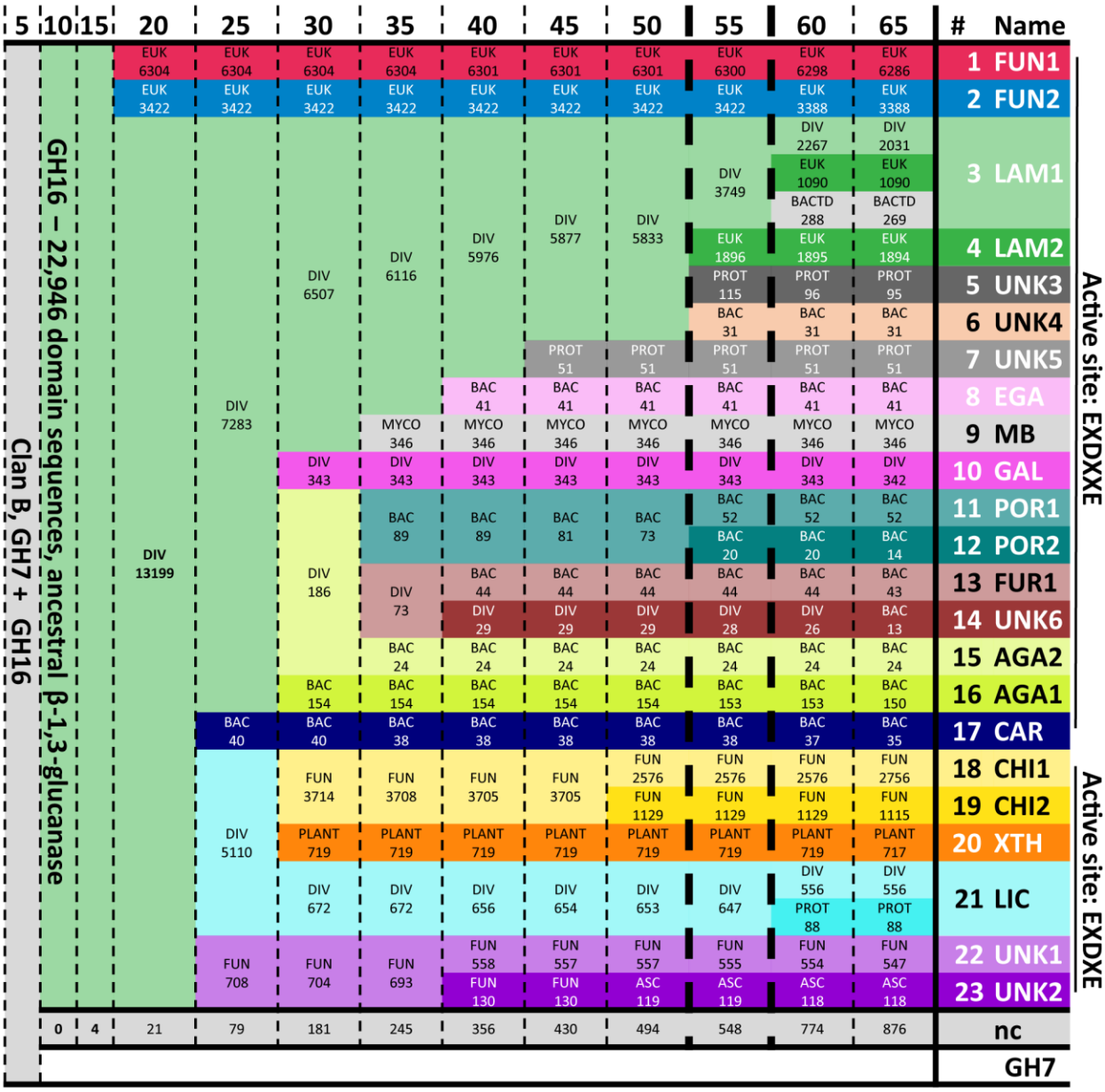

Figure 2. Summary of GH16 Sequence Similarity Networks. Summary of the subfamilies created in Sequence Similarity Networks (SSNs) under thresholds from $E = 10^{-5}$ to 10^{-65}. The top row indicates the SSN clustering threshold defining each column (*e.g.* "35" corresponds to an *E*-value threshold of 10^{-35}). Rows represent the emergent subfamilies (colored individually) and their stability across thresholds. Labels in the subfamilies indicate the number of sequence members as well as the taxonomic range (ASC: Ascomyocota, BAC: Bacteria; BACTD: Bacteroidetes; DIV: Multiple kingdoms, EUK: Eukaryota; FUN: Fungi; MYCO: *Mycobacterium*; PLANT: Plantae; PROT: Proteobacteria). Definitive subfamilies defined based on the $E = 10^{-55}$ threshold (column marked with bold dashed lines) are numbered in the right-most column, in ascending order according to the family size/sequence members. Subfamily mnemonics assigned based on known activities or taxonomic distribution are as follows: AGA: β-agarases; CAR: κ-carrageenase; CHI: chitin β(1,6)-glucanosyltransferase; EGA: *endo*-β(1,4)-galactosidases; FUN: Fungal; FUR: Furcellaranase; GAL: *endo*-β(1,3)-galactanases; LAM: *endo*-β-glucanases; LIC: *endo*-β(1,3)/β(1,4)-glucanase MB: Mycobacterium; POR: β-porphyranases; UNK: Unknown; XTH: Xyloglucan *endo*-

tranglycosylase / *endo*-hydrolase). The bottom row show the non-classified (nc) sequences, not assigned to any subfamily (548 of 22946 total GH16 sequences at the 10^{-55} threshold).

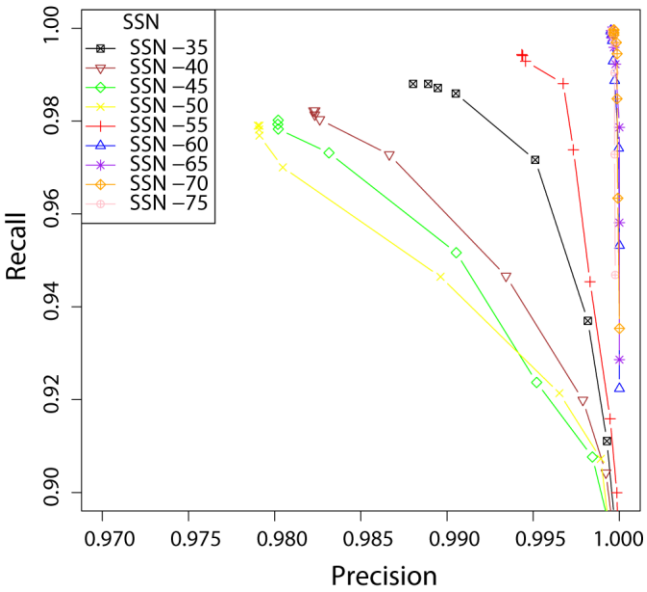

Figure 3. Performance of GH16 Hidden Markov Model libraries. HMM libraries of GH16 subfamilies, generated from the SSN at each threshold (color-coded in the legend), were evaluated in their ability to assign each GH16 module to the correct subfamily delineated by the individual SSNs. The curves show the evolution of the precision and recall (see Methods for definitions) with increasing SSN *E*-value cutoff (*cf.* Figure 2 and Figure 4), with points corresponding to variation in HMM *E*-value thresholds.

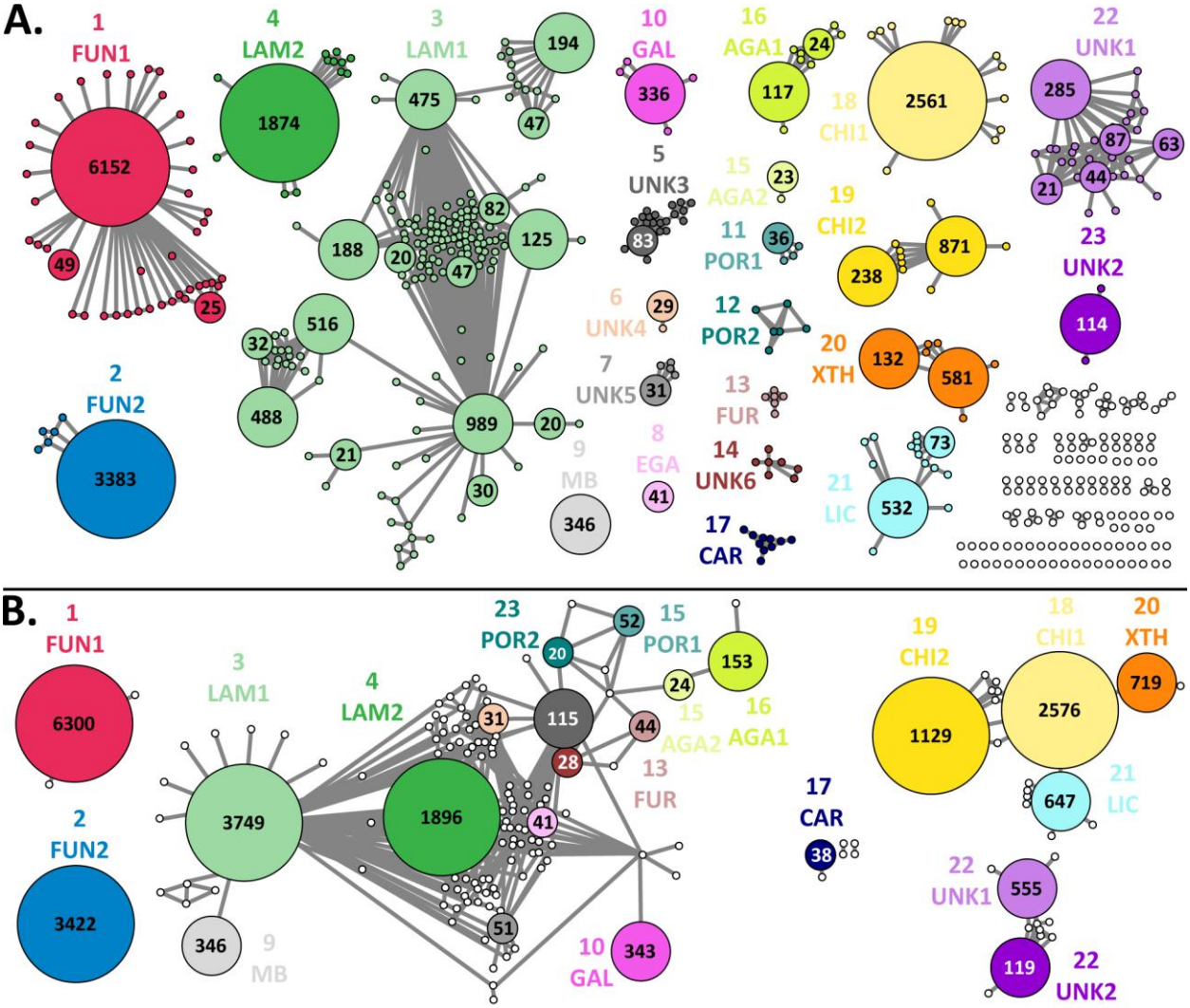

Figure 4. Sequence Similarity Networks of 22,946 GH16 sequences. a. Edges represent an *E*-value threshold below 10^{-55}. Meta-nodes represent highly similar sequences ($E > 10^{-85}$); only meta-nodes containing 20 or more sequences are enlarged, with the number of merged sequences indicated. The network defines 23 subfamilies (see Fig. 2 for subfamily numbering and mnemonics). Clusters that lack sufficient taxonomic diversity or size to define subfamilies are indicated in white. **b.** Edges represent an *E*-value threshold below 10^{-25}. Meta-nodes represent defined subfamilies in A. ($E > 10^{-55}$); the network displays the basic relationship of subfamilies at this relaxed threshold (*cf.* Figure 2).

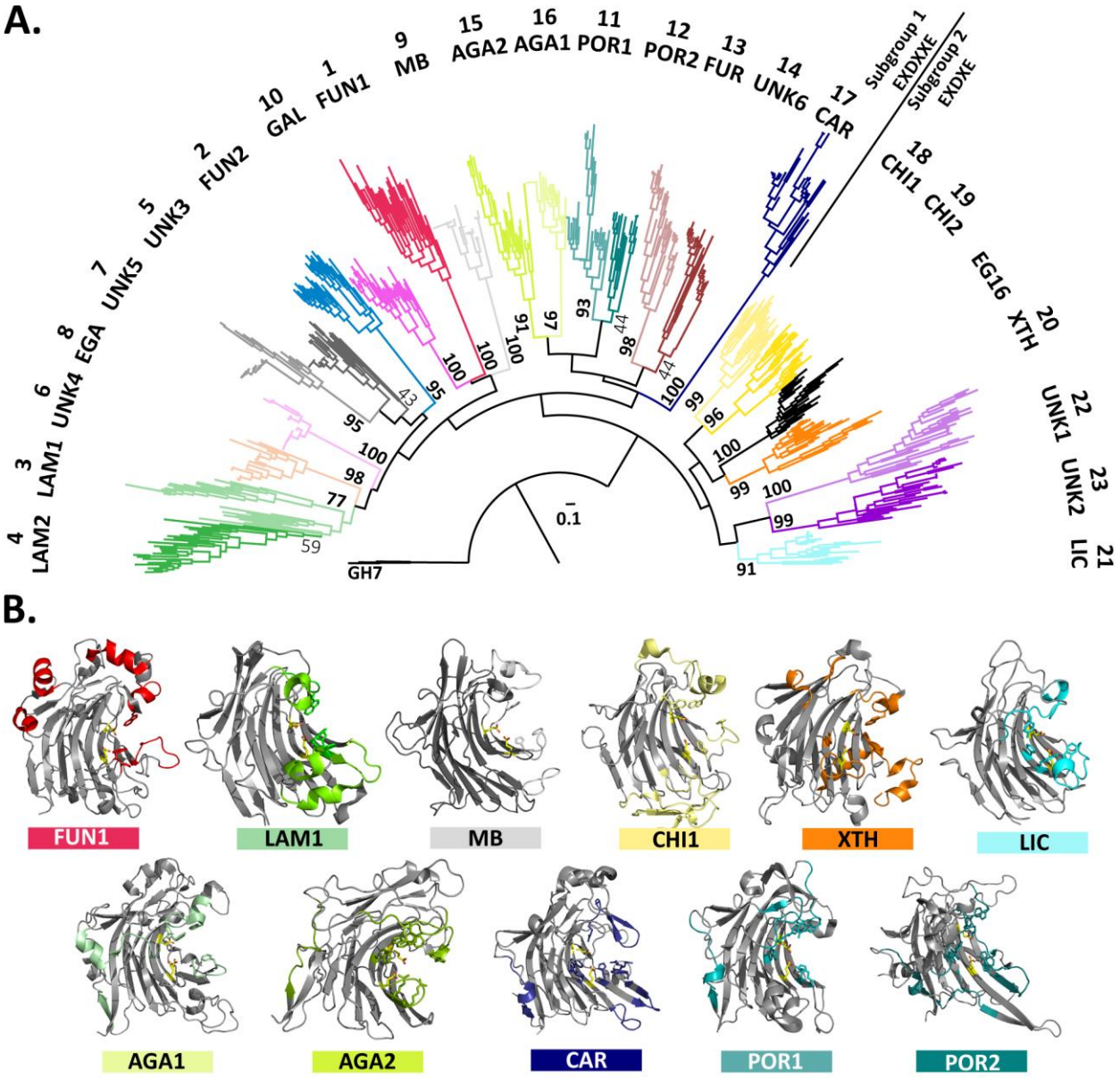

Figure 5. Phylogenetic tree and structure-function relationships of GH16. a. Maximum-likelihood phylogenetic tree was generated using up to 30 representative sequences for each GH16 subfamily defined by the Sequence Similarity Network shown in Fig. 4. Three GH7 cellulases (GH7 and GH16 constitute clan GH-B (7)) were used to root the tree. Bootstrap values based on 100 replicates are shown. The tree separates (indicated by a line) GH16 enzymes with the β-bulge active-site motif EXDXXE from those with the β-strand active-site motif EXDXE, concordant with previous analyses (30, 42). Branch coloring is identical to that used in Figs. 2 and 4; subfamily numbering and mnemonics are given in Fig. 2. Note that although the GH16 sequences were randomly selected to capture subfamily diversity in an unbiased manner, without regard to functionally or structurally characterized members, subfamily membership of all GH16 members is available on the actively-curated CAZy database via URL http://www.cazy.org/GH16.html **b.** Ribbon drawings of 3D structures of representative subfamily members (where present, see Table 1). Loops, structural elements and residues that are characteristic of a given subfamily are colored with their respective color (color-bar underneath the structural icon), the same as in the phylogenetic tree in A.