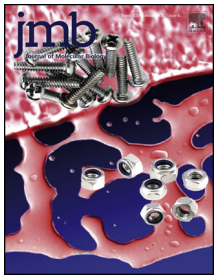

Extent and Origins of Functional Diversity in a Subfamily of Glycoside Hydrolases

Evan M. Glasgow[1,2,†], Kirk A. Vander Meulen[1,2,†], Taichi E. Takasuka[1,2,3], Christopher M. Bianchetti[1,2,4], Lai F. Bergeman[1,2], Samuel Deutsch[5] and Brian G. Fox[1,2]

*1 - **Great Lakes Bioenergy Research Center,** Madison, WI 53706 USA*
*2 - **Department of Biochemistry,** University of Wisconsin, Madison, WI 53706 USA*
*3 - **Research Faculty of Agriculture,** Hokkaido University, Sapporo, 060-8589 Japan*
*4 - **Department of Chemistry,** University of Wisconsin, Oshkosh, 54901 USA*
*5 - **DOE Joint Genome Institute,** Walnut Creek, CA 94598 USA*

***Correspondence to Brian G. Fox:** Department of Biochemistry, University of Wisconsin–Madison, 141B Hector F. DeLuca Biochemistry Laboratories, 433 Babcock Drive, Madison, WI 54706-1544, USA. bgfox@wisc.edu*
https://doi.org/10.1016/j.jmb.2019.01.024
Edited by Dan Tawfik

Abstract

Some glycoside hydrolases have broad specificity for hydrolysis of glycosidic bonds, potentially increasing their functional utility and flexibility in physiological and industrial applications. To deepen the understanding of the structural and evolutionary driving forces underlying specificity patterns in glycoside hydrolase family 5, we quantitatively screened the activity of the catalytic core domains from subfamily 4 (GH5_4) and closely related enzymes on four substrates: lichenan, xylan, mannan, and xyloglucan. Phylogenetic analysis revealed that GH5_4 consists of three major clades, and one of these clades, referred to here as clade 3, displayed average specific activities of 4.2 and 1.2 U/mg on lichenan and xylan, approximately 1 order of magnitude larger than the average for active enzymes in clades 1 and 2. Enzymes in clade 3 also more consistently met assay detection thresholds for reaction with all four substrates. We also identified a subfamily-wide positive correlation between lichenase and xylanase activities, as well as a weaker relationship between lichenase and xyloglucanase. To connect these results to structural features, we used the structure of CelE from *Hungateiclostridium thermocellum* (PDB 4IM4) as an example clade 3 enzyme with activities on all four substrates. Comparison of the sequence and structure of this enzyme with others throughout GH5_4 and neighboring subfamilies reveals at least two residues (H149 and W203) that are linked to strong activity across the substrates. Placing GH5_4 in context with other related subfamilies, we highlight several possibilities for the ongoing evolutionary specialization of GH5_4 enzymes.

Introduction

Enzymes have acquired extraordinary breadth of substrate specificity over millions of years of mutation and selection of ancestral precursors [1–3]. One common and frequently recurring protein fold, the $(\beta/\alpha)_8$ barrel (also known as a TIM barrel, first recognized in triose phosphate isomerase [4]), has evolved to span perhaps the widest range of catalytic functions among all folding motifs, although it is most frequently observed to serve as a hydrolase [5–7]. TIM barrel enzymes possess a well-conserved core and malleable active site pocket whose function is easily rationalized to be highly adaptable.

Glycoside hydrolase family 5 (GH5) is an expansive TIM barrel GH family, with over 13,000 listings at present in the Carbohydrate-Active enZyme (CAZy) database [8]. GH5 enzymes have been implicated in the hydrolysis of a diverse set of oligo- and polysaccharide substrates, with nearly 20 enzyme activity categories identified [9]. GH5 is itself part of a larger "clan" referred to as GH-A, with all member families sharing two conserved glutamate residues that are used for hydrolysis of glycosidic bonds. One glutamate, located at the end of β-strand 7, carries out nucleophilic attack on the anomeric carbon, while a second glutamate at the end of β-strand 4 activates water for attack on the covalent glycosyl-enzyme

intermediate, which hydrolyzes the glycosidic bond with stereochemical retention [10,11].

While substrate specificity is often a defining feature in enzyme-catalyzed processes, enzymes with more relaxed specificity are valuable tools in textiles, materials and food processing, pharmaceuticals, detergents, paper and pulp industries, and energy production [12–14]. In biofuels applications, GHs with broad specificity can potentially hydrolyze many different types of plant cell wall polysaccharides [15,16], offering the potential for simplification of the complex enzyme preparations used to produce fermentable sugars from plant biomass [17]. This work focuses on understanding the extent and structural origins of this broad substrate specificity.

Broad substrate specificity has been noted in several GH5 subfamilies. In GH5 subfamily 4 (GH5_4), both β-1,4-glucanase and xylanase activity have often been observed [18–25]. Another study from the closely related subfamily GH5_25 noted the enzymes' ability to hydrolyze both glucan and mannan [26]. In the case of the multidomain enzyme CelE from *Hungateiclostridium thermocellum*, we demonstrated the ability of its GH5_4 domain to hydrolyze the β-(1,4) backbone for three types of polysaccharides: cellulose, xylan, and mannan. We therefore referred to it as a "CMX" enzyme [27] and solved its crystal structure (PDB 4IM4; Bianchetti, C.M., Takasuka, T.E., Fox, B.G., to be published). In addition, CelE can hydrolyze branched and mixed-sugar glycans (e.g., xyloglucans) [28]. Indeed, some researchers have considered xyloglucanase activity to be a distinguishing characteristic of the GH5_4 subfamily [9,29–34].

Our emphasis in this work is in rationalizing and predicting the spectrum of catalytic versatility, in terms of polysaccharide backbone hydrolysis, that has been observed in GH5_4. To do this, we used gene synthesis and cell-free protein expression to perform a quantitative biochemical activity screen across the entire subfamily. This systematic mapping of function to phylogeny revealed clade-level substrate specificity profiles, with some clades significantly outperforming others on the unbranched substrates lichenan, mannan, and xylan. Furthermore, activities on lichenan and xylan are observed to correlate throughout the subfamily. The evolutionary implications and key sequence and structural features associated with this reactivity are discussed.

Results

Subfamily-wide enzyme activity screen

DNA synthesis capabilities at the Joint Genome Institute were combined with small-scale cell-free protein production to express the catalytic domain of 243 GH5 enzymes. Two hundred thirty-eight of these enzymes belong to GH5_4 and were selected to span the sequence diversity within the subfamily; five additional sequences from related GH5 subfamilies were also included. We first assayed each enzyme on lichenan, beechwood xylan, and β-1,4-mannan to assess the scope of activity on unbranched, homopolymeric substrates. We used lichenan, a mixed β-1,3 and β-1,4 linked glucan, in place of cellulose for measuring endoglucanase activity because of its increased solubility and ease of use in low-volume plate-based assays. Hydrolysis by some GH5 endoglucanases has been shown to occur at both β-1,3 and β-1,4 linkages with similar kinetics, provided a 1,4 linkage is present between the −2 and −1 subsites [21,22,32]. While it is not possible to determine which linkages (β-1,3 or β-1,4) are being hydrolyzed in our experimental setup, our goal was to maximize the likelihood of activity detection and increase the signal-to-noise ratio of the activity data. Beechwood xylan was used for endoxylanase activity screening because of its relatively high solubility and low degree of substitution. Finally, we avoided using more soluble alternatives to linear β-1,4-mannan (such as glucomannan) or carboxymethylcellulose because substrates with mixed-sugar backbones or other chemical derivatizations might confound the assignment of specificity.

Assays for each enzyme–substrate pair were performed over a wide range of temperature (30, 50, and 70 °C) and pH (4, 5, 6, 7, and 8), and the maximum specific activities across these conditions were determined. The rationale behind this approach was to increase the probability of identifying an activity above the detection threshold and to allow subsequent comparison of each activity near its optimum. Using this approach, at least one activity was measurable for 229 of the 243 enzymes.

Activities were most commonly detected on lichenan, for which 226 enzymes demonstrated measurable activity. Xylanase activity was the second most common, at 204. Mannanase activity was the least commonly observed although still detected in over half (139) of the enzymes. A Venn diagram representation of these threshold data paints a hierarchical picture for the specificity profiles (Fig. 1A). With only one exception, enzymes with xylanase activity are a subset of those with lichenase activity. Less consistently but similarly, mannanase activity generally occurs in the subset of enzymes possessing the other two activities.

The optimum temperature and pH determined across screen conditions are illustrated by histograms in Fig. 1B. Activities on both lichenan and xylan are usually optimal at either 50 °C (60% and 61% of detected enzymes, respectively) or 30 °C (35% and 34%, respectively). In contrast, mannanase activity is almost always optimal at 30 °C (91%). pH optima display a similar contrast between lichenase and xylanase activities relative to mannanase activities. Lichenase and xylanase activities are usually maximal at pH 6 (56% for lichenase measurements, 62% for xylanase),

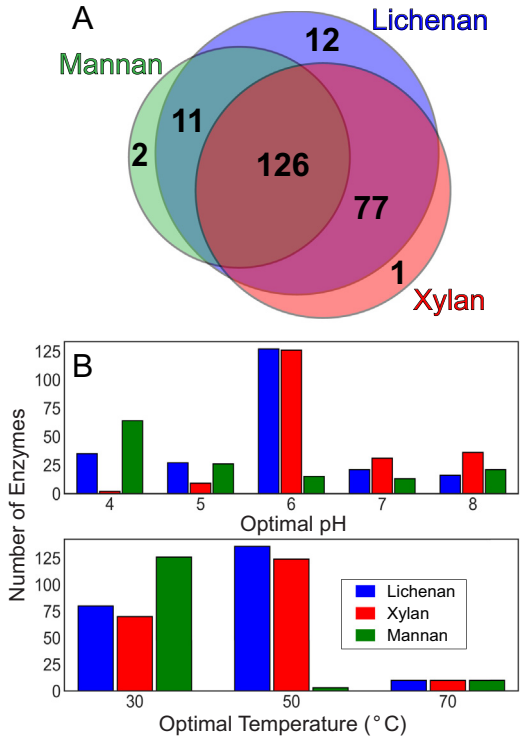

Fig. 1. Activity data characteristics. Blue, red, and green represent enzyme performance on lichenan, xylan, and mannan, respectively. (A) Venn diagram showing assay-defined mono-, bi-, and trifunctional phenotypes. (B) Histograms of optimal pH (top panel) and temperature (lower panel) among enzymes with detected activities.

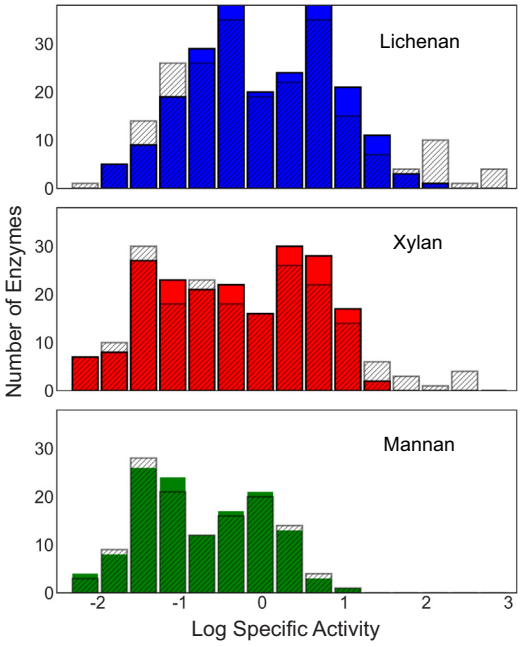

Fig. 2. Histogram of $\log_{10}$ specific activities for GH5_4 enzymes on lichenan (top panel), xylan (middle), and mannan (lower); bin size = 0.36 log unit. For each panel, solid bars describe raw specific activity values, and open-hatched bars describe the histogram for specific activities corrected for substrate depletion. Measurements that likely indicated lower limits were counted in the top bin for each substrate (3 such examples for lichenan and 4 for xylan).

while mannanase activities are most commonly optimal at pH 4 (46%). There is no particular pH and/or temperature condition where the maximum specific activities are significantly higher than for others.

The screen-wide specific activity values span at least 4 orders of magnitude (roughly 5×10^{-3} U/mg to slightly greater than 1×10^{1} U/mg; see Materials and Methods). Log specific activity values on each substrate display a bimodal distribution (Fig. 2). For lichenan, two clear peaks occur in the histogram with bins centered at specific activity values of 0.42 and 4.9 U/mg. The number of enzymes with activity values at the lower end of the distribution increases stepwise starting from zero in the lowest bin, suggesting that the sensitivity of the method largely captures the distribution of activities. The upper end of the activity distribution drops more sharply, possibly due to an underestimation of the highest specific activity values caused by substrate depletion in these single-measurement calculations. Figure 2 overlays the histogram resulting from naïve specific activity calculations (activity divided by time) with a histogram using specific activities that have been adjusted to a specific activity estimated at 5% substrate depletion (open-hatched bars, see Materials and Methods). The most obvious effect of this correction on the histogram is to broaden the top end (right-hand side) of the distribution.

The xylanase data resemble the bimodal pattern of the lichenase histogram, but with a slight shift toward lower log specific activities. Bins with maximum occupancy for the two modes are centered at 0.036 and 2.2 U/mg. Consistent with this, the irregular shape at the lower end of the xylanase histogram suggests that some specific activities lie below the screen's detection threshold. At the upper end of the histogram, which does not extend as far as lichenan, some values are still likely underestimates as indicated by specific activity estimated at 5% substrate depletion (open-hatched bars). The mannanase histogram appears further leftward-shifted and clipped at the lower end. At the upper end, although the bin sizes decrease somewhat abruptly to 3 and 1 for the bins centered at 4.9 and 11.0 U/mg, these maximum activities are well within the experimental range, suggesting no underestimation for mannanase measurements.

Lichenase and xylanase activities are highly correlated

Figure 3 plots the respective log specific activities for each enzyme on lichenan against xylan; a linear least-squares fit yields a slope of 0.61 with R^2 of 0.52. The

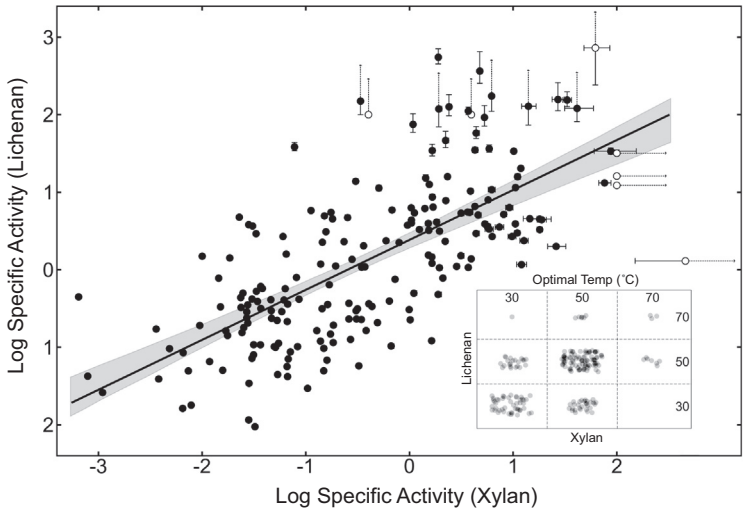

Fig. 3. Correlation in enzyme activities on lichenan and xylan. (Main figure) Yield-normalized log specific activities (see Materials and Methods) for each enzyme on lichenan are plotted against corresponding activity on xylan, for all enzymes possessing both activities. Error bars reflect the impact of ±5% uncertainty in the measured yield on the plotted specific activity. Open symbols indicate that the measurement reflects a lower bound, and a dashed arrow indicates that an upper bound cannot be determined. The solid line and surrounding shaded region illustrate the best-fit line and the corresponding resampling-determined 90% confidence interval. (Inset) Corresponding plot of optimal temperatures for lichenase and xylanase activities. Transparency and jitter are added to each point to aid in visualization of tallies for each bin.

figure and analysis only include enzymes with detectable activities on both substrates; consistent with the correlation, the 22 lichenase enzymes with undetectable xylanase activity exhibit significantly lower lichenase activities (mean 0.03 U/mg) than do enzymes with both activities (mean 1.4 U/mg). The inset demonstrates that the correlation extends to the temperature optima for each enzyme; 138 out of 203 enzymes with both activities share optimal lichenase and xylanase temperatures. A pH relationship also exists: in general, an enzyme's pH optimum for xylanase activity is higher than or equal to its optimum for lichenase (not shown).

Relationships to mannanase activity are weaker. There is no significant subfamily-wide correlation between activities on mannan and either lichenan or xylan. Also, while average activities for both lichenase (1.6 U/mg *versus* 0.5 U/mg) and xylanase (0.5 U/mg *versus* 0.3 U/mg) are slightly higher for the pool of enzymes that also possess mannanase activity, these 2- and 3-fold differences are less compelling than the 50-fold differences observed for lichenases with (1.4 U/mg) or without (0.03 U/mg) xylanase activity noted above. On the other hand, there do appear to be exceptions, suggesting that increased mannanase activity is not intrinsically antagonistic with lichenase or xylanase activity (clades 1B, 2B, and 3B; see below).

Xyloglucanase activity

A subset (110) of enzymes were also assayed on xyloglucan, and the activity is observed to be widely distributed across GH5_4. Of the 110 enzymes tested, 98 had measurable xyloglucanase activity, with specific activities spanning at least 4 orders of magnitude (Supplemental Data). All seven enzymes in this subset with no detectable activity on lichenan, mannan, or xylan showed xyloglucanase activity (up to ~5 U/mg), and a weak correlation was measurable between xyloglucanase and lichenase activities (Supplemental Data).

Evolution of GH5_4 enzymes

A Bayesian inference phylogenetic tree was constructed from the catalytic core sequences for 638 GH5_4 and closely related enzymes (Fig. 4). All but one of the sequences in this study previously annotated in the CAZy database to be part of subfamily 4 lie within three major clades; henceforth, these are referred to as GH5_4 clades 1, 2, and 3. Clade 1 diverged from the rest of GH5_4 first, in a major topological event that also resulted in significant differences from the rest of the subfamily. Clade 1 enzymes universally and exclusively possess a catalytic core-augmenting Ig-like-CBM46 fusion module [33,34] (referred to as CBM_X2 in the Pfam database), indicating that it was either added or optimized along the clade 1-specific ancestral lineage. In addition, it appears that at some point in the clade 2/3-specific branch, an ancestral enzyme was incorporated by a rumen organism, further differentiating the evolutionary pathways (see Supplemental Data).

The three major clades are denoted in the figure through yellow, brown, and orange highlighting, respectively; subcategorizations into 1A–1B, 2A–2D, and 3A–3G are indicated by the varied shading within the clade. Previously reported trees had similar structure, with the major exception being that they did not have sufficient support to assign clades 1A and 1B within a larger major GH5_4 clade [9,26,32]. The subfamily 4-focused tree constructed here reports a clade 1 root node with high confidence (Supplemental Data), and this is corroborated by the exclusivity of the CBM46-containing gene construct within the clade.

Three sequences from the closely related subfamilies 25 and 37 were also included in the experimental data set, as well as two sequences from the more

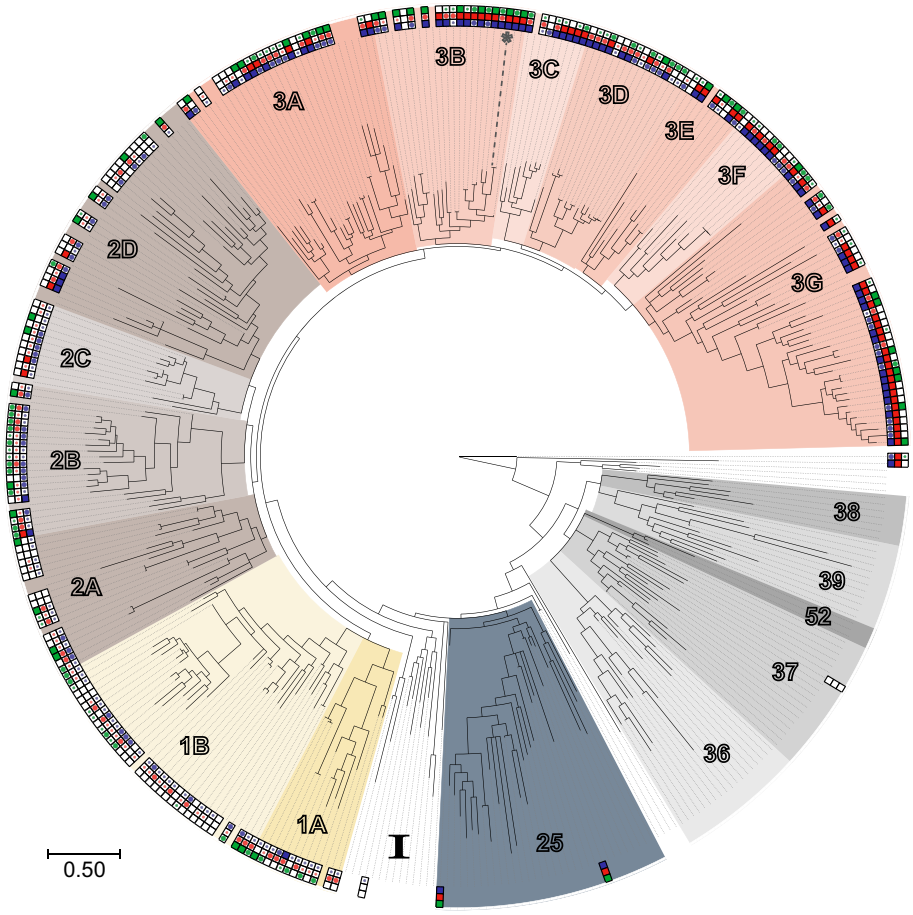

Fig. 4. Consensus phylogram constructed using catalytic core sequences for GH5 subfamilies 4, 25, 36, 37, 38, 39, and 52. Nodes with <50% support are displayed as polytomies. Major clades within GH5_4 are denoted by yellow, brown, and orange highlighting, with varying saturation for each corresponding to subclade partitioning; other subfamilies are denoted by varying shades of gray. Enzyme sequences that were tested for activities are indicated by a three-cell table at the outer edge of the tree. The presence of blue, red, or green circles within each cell indicates lichenase, xylanase, or mannanase activity, respectively, and circle size indicates activity strength. The largest, most intense symbols reflect enzymes with specific activities ranking in the top third. The location of CelE is also highlighted (*).

distantly related subfamilies 2 and 5. The tree contextualizes these enzymes in the evolution of GH5_4 and its closely related subfamilies (25, 36, 37, 38, 39, and 52). Subfamilies 2 and 5 are more distantly related to GH5_4 [9], so the GH5_2 enzyme was set as the tree root-defining outgroup enzyme, located at the right-hand side of the figure. The GH5_5 enzyme appears directly adjacent.

Phylogenetic ranking and analysis of activities

Experimentally determined lichenase-, xylanase-, and mannanase-specific activities are mapped to the exterior of the tree in Fig. 4. Several clusters of high and low activities are evident, indicating the utility of grouping and studying enzymes according to phylogeny. Averages within clades and subclades are listed in Table 1. In this section, major generalizations are described.

The first and perhaps most obvious generalization is that clade 3 possesses the best lichenase and xylanase enzymes. The average clade 3-specific activities on lichenan (4.24 U/mg) and xylan (1.17 U/mg) exceed those of clades 1 and 2 by approximately an order of magnitude (Table 1). This clade-based segmentation underlies the bimodal histogram in Fig. 2B (this also appears to be the case for xyloglucanase activities; see Supplemental Data).

The second major point is that, while clades 1 and 2 lag in performance relative to clade 3, the average lichenase- and xylanase-specific activities in clade 2 (0.36 and 0.11 U/mg) are slightly greater than those in clade 1 (0.13 and 0.06 U/mg). It should be noted here that clade 1 enzymes may be intrinsically

Table 1. Average specific activity values

Clade	Seqs tested[a] (coverage [b])	Lichenan		Xylan		Mannan	
		%Active[c]	Spec. activity (μmol min^{-1} mg^{-1})[d]	%Active[c]	Spec. activity (μmol min^{-1} mg^{-1})[d]	%Active[c]	Spec. activity (μmol min^{-1} mg^{-1})[d]
All (1–3)	237 (57%)	93 (84)	1.09 (0.92–1.30)	84 (62)	0.35 (0.29–0.42)	57 (37)	0.12 (0.10–0.14)
1	46 (51%)	87 (57)	0.13 (0.10–0.16)	76 (22)	0.06 (0.05–0.07)	43 (20)	0.11 (0.08–0.15)
1A	12 (64%)	100 (50)	0.16 (0.10–0.24)	83 (25)	0.05 (0.03–0.07)	50 (25)	0.10 (0.06–0.15)
1B	34 (46%)	82 (59)	0.12 (0.09–0.15)	74 (21)	0.06 (0.05–0.08)	41 (18)	0.12 (0.08–0.18)
2	69 (52%)	90 (84)	0.36 (0.29–0.44)	80 (51)	0.11 (0.09–0.14)	48 (36)	0.16 (0.11–0.22)
2A	17 (66%)	82 (65)	0.15 (0.11–0.20)	65 (35)	0.07 (0.05–0.10)	41 (18)	0.17 (0.08–0.39)
2B	20 (55%)	100 (100)	0.59 (0.45–0.77)	95 (75)	0.13 (0.09–0.18)	85 (80)	0.14 (0.09–0.20)
2C	11 (37%)	100 (100)	0.19 (0.15–0.25)	91 (55)	0.10 (0.05–0.20)	18 (9)	0.45 (0.40–0.45)
2D	21 (47%)	81 (76)	0.46 (0.27–0.80)[e]	71 (38)	0.12 (0.07–0.20)	33 (24)	0.18 (0.11–0.31)
3	122 (64%)	98 (95)	4.24 (3.36–5.40)	90 (84)	1.17 (0.90–1.55)	68 (43)	0.11 (0.09–0.13)
3A	23 (51%)	100 (87)	6.15 (3.04–12.23)[e]	83 (83)	0.25 (0.16–0.38)	70 (61)	0.14 (0.08–0.26)
3B	19 (53%)	95 (95)	6.39 (4.16–9.86)[e]	95 (84)	1.09 (0.54–2.26)	89 (42)	0.66 (0.45–0.94)
3C	8 (63%)	100 (100)	2.49 (1.32–4.61)	88 (88)	1.25 (0.83–1.93)	75 (63)	0.06 (0.04–0.09)
3D	16 (100%)	100 (100)	2.32 (1.58–3.46)	88 (75)	0.35 (0.23–0.53)	69 (38)	0.02 (0.01–0.05)
3E	8 (79%)	75 (75)	4.83 (1.68–12.68)	75 (88)	6.02 (1.20–31.76)[e]	100 (75)	0.07 (0.04–0.12)
3F	10 (83%)	100 (100)	18.99 (7.61–45.38)[e]	80 (70)	1.38 (0.94–2.02)	50 (50)	0.02 (0.01–0.03)
3G	38 (64%)	100 (100)	2.73 (1.97–3.79)[e]	100 (92)	1.99 (1.19–3.26)[e]	53 (24)	0.13 (0.08–0.19)

[a] Number of sequences tested from each group.
[b] Percent of tree length for each clade represented by the included sequences.
[c] Percent with specific activities >0; in parentheses, percent that also satisfy high-stringency threshold for DNS signal intensity relative to an experimentally determined distribution of substrate-only measurements ($p < 0.005$).
[d] Average and 90% confidence intervals, determined by bootstrap.
[e] One or more values may be lower limits, so tabulated values may be underestimates.

disadvantaged by comparison of the reactivity of the catalytic cores alone because their co-evolved CBM46 module has been omitted [33,34].

The third point from Table 1 is that the screen-wide average specific activity for mannan (0.12 U/mg) is significantly lower than for either lichenan (1.09 U/mg) or xylan (0.35 U/mg); this is also apparent from the histograms in Fig. 2. The cause may be limited substrate accessibility due to the crystalline nature of pure β-1,4 mannan [35], which is supported by the relatively weaker reactivity observed with PASC relative to lichenan (Supplemental Data). Regardless of the root cause, implications for this study are twofold. First, signal-to-noise is lower for the mannanase experiments. Second, comparison of mannanase activity among groups of enzymes must account for the fact that only the most active enzymes from a spectrum of mannanase activities are quantified.

For activity on mannan, we consider that the most reliable metric of average enzyme performance is simply the number of threshold-exceeding enzymes, in other words, screen "hits." Here again, clade 3 ranks at the top, with 68% of its members showing detectable activity, trailed by clade 2 (48%) and clade 1 (43%).

Phylogenetic subranking and analysis of activities

Increasing the granularity of phylogenetic groupings makes possible a number of additional observations. Specific activity data for individual enzymes reacted with lichenan and xylan are replotted in Fig. 5, with an emphasis on clade membership (see also Table 1). Representatives from clades 1 and 2 are clustered toward the lower left of the plot, indicating that most enzymes from these groups possess diminished lichenase- and xylanase-specific activities relative to clade 3. Enzymes assayed from related subfamilies (25, 2, and 5) show higher specific activities similar to clade 3.

A correlation between lichenase and xylanase activities persists within each of the major clades, although it is most compelling in clade 3. Table 2 lists the results for all identified statistically relevant log specific activity relationships obtained from the full 243 enzyme data set and shows that this relationship is also significant within a number of subclades. Relationships between either activity and mannanase are less convincing, although compelling correlations are observed between lichenase and mannanase within subclades 1B and 2B.

Within clade 3, members of clade 3B, exemplified by *H. thermocellum* CelE and predominantly arising from free-living Clostridia, might be the best overall performers within the assay conditions of this study. This group possesses the second and fifth highest average lichenase- and xylanase-specific activities, as well as the second largest fraction of detected activities on mannan. The fungal-dominated clade 3D is at the other extreme and ranks last in specific activity for lichenan and xylan and third lowest in mannanase as detected by screen hits.

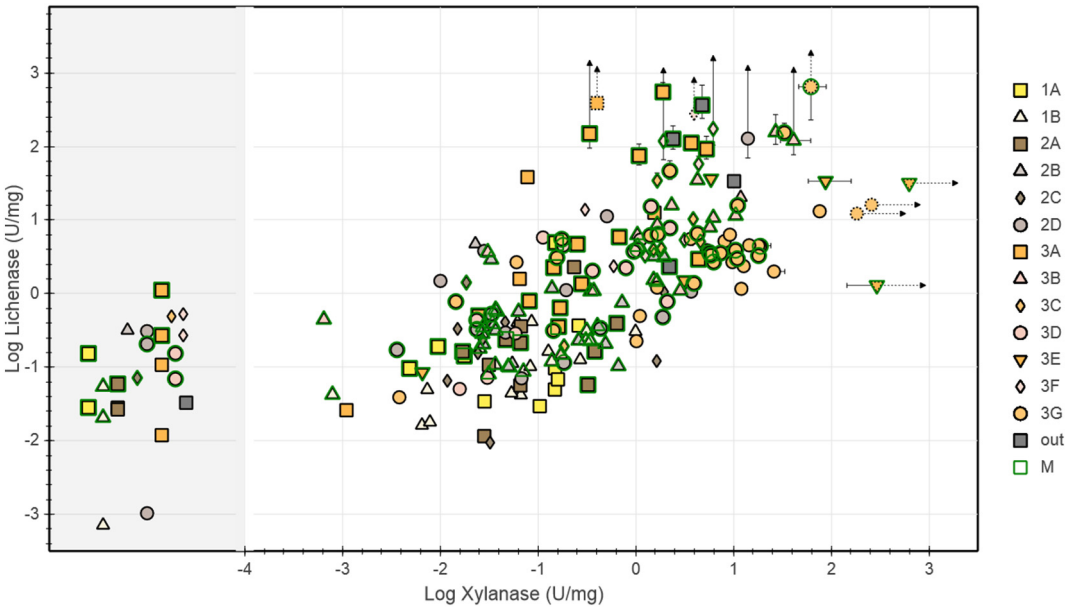

Fig. 5. Phylogenetic grouping-focused plots of base-10 logarithm lichenase *versus* xylanase activities. Symbol colors and types reflect clade membership as indicated, with "out" indicating outgroup enzymes from related subfamilies; green outlines (M) highlight enzymes for which mannanase activity was also detected. Error bars report uncertainty in plotted specific activities caused by propagation of uncertainty in yield measurements, as in Fig. 3. For clarity, only uncertainties of this type greater than 0.1 log units are plotted. Specific activity measurements that are likely lower bounds in either lichenase or xylanase values are denoted by symbols with dashed borders. Arrowheads on error bars similarly indicate that an upper limit uncertainty value cannot be determined. Scatter points in the (left-hand) off-scale shaded region exhibited detectable lichenase, but not xylanase activity. Interactive plot features allow inspection of accession code, source organism, measured specific activities, and optimum conditions for each enzyme. Interactive plot features allow inspection of accession codes, source organism, measured specific activities, and optimum conditions for each enzyme, as well as selective viewing of individual clades by clicking the corresponding legend entry Specific activities are in U/mg (n.d. = not detectable), optimum temperatures corresponding to reported specific activity measurements are in °C for lichenase, xylanase, and mannanase, respectively.. Xyloglucanase-specific activity measurements were obtained in a separate set of experiments, with all data collected at 30 °C.

Table 2. Statistically significant[a] linear regression results

Clade	Data points	ρ[a,b]	p[a]	Slope[b]
Lichenase–xylanase				$\partial\log L/\partial\log X$
All data	203	0.71 (0.66–0.76)	$< 1 \times 10^{-31}$	0.6 (0.5–0.7)
Clades 1–3	199	0.71 (0.65–0.76)	$< 1 \times 10^{-30}$	0.6 (0.4–0.7)
Clade 1	35	0.43 (0.16–0.61)	0.01	0.3 (0.2–0.5)
Clade 1B	25	0.47 (0.11–0.71)	0.02	0.4 (0.2–0.5)
Clade 2	54	0.21 (−0.04–0.43)	0.12	0.4 (0.1–0.7)
Clade 2A	11	0.50 (−0.20–0.78)	0.12	0.4 (−0.1–1.1)
Clade 3	110	0.50 (0.35–0.61)	$< 1 \times 10^{-7}$	0.5 (0.2–0.5)
Clade 3A	19	0.69 (0.40–0.84)	< 0.01	0.9 (0.5–1.3)
Clade 3B	18	0.83 (0.58–0.93)	$< 1 \times 10^{-4}$	0.5 (0.3–1.2)
Clade 3D	14	0.68 (0.27–0.92)	0.01	0.8 (0.4–1.1)
Clade 3F	8	0.67 (0.09–0.93)	0.07	1.2 (0.4–4.0)
Clade 3G	38	0.49 (0.18–0.69)	< 0.01	0.4 (0.2–0.5)
Mannanase–xylanase				$\partial\log M/\partial\log X$
Clade 3A	14	0.41 (−0.02–0.71)	0.14	0.6 (0.1–1.4)
Clade 3F	5	0.84 (0.11–1.0)	0.07	2.3 (0.1–3.7)
Mannanase–lichenase				$\partial\log M/\partial\log L$
Clade 1B	14	0.45 (0.06–0.70)	0.11	0.9 (0.4–1.8)
Clade 2B	17	0.54 (0.23–0.76)	0.02	0.7 (0.4–1.1)

[a] All activity relationships and Spearman's rank coefficient ρ for which p value is <0.2.
[b] Best estimate and 90% confidence interval.

Two clade 3 subgroups show statistically significant deviation from the lichenan- *versus* xylan-specific activity trendline shown in Fig. 3, indicating a specialization toward either lichenase or xylanase reactivity relative to the other enzymes. Clade 3A enzymes behave as relative lichenan specialists, thus comprising the first group, exhibiting a distribution in offsets from the trendline that is statistically distinct from the rest of clade 3 (two-way ANOVA, $p < 0.001$), as well as several notably large specific activity values on lichenan (Fig. 5). In contrast, clade 3G activities are skewed toward xylanase (i.e., below the trendline; two-way ANOVA, $p < 0.001$). Overall, an improvement in xylanase activity is observed in the sub-branch composed of clades 3E, 3F, and 3G: enzymes from these clades exhibit the three highest average xylanase-specific activities.

The most interesting exception to the typically lowered activities in clades 1 and 2 is the performance on linear substrates in clade 2B, a subclade dominated by rumen bacteria that possesses (within clades 1 and 2) the highest average specific activities on both lichenan and xylan. In addition, a remarkable 85% of enzymes in this subclade showed mannanase activity, significantly higher than all other subclades in clades 1 and 2, as well as most in clade 3. In fact, the top four mannanase subclades in terms of percentage of tested enzymes above the minimum dynamic range are, in order, clade 3E, 3B, 2B, and 2C (see Table 1).

Discussion

Endoglucanases and multifunctionality

The ability of a single enzyme to hydrolyze multiple substrates potentially increases its utility to microorganisms and provides an opportunity to simplify enzyme preparations used in industrial applications. Both xylan and lichenan are long, unbranched chains of pyranose sugars composed entirely (xylan) or mostly (lichenan) of β-1,4 linkages, differing primarily by the presence of an equatorial C6 hydroxymethyl group. Thus, the ability of an endoglucanase featuring an open binding cleft to hydrolyze β-1,4 glycosidic bonds in both substrates is not altogether surprising. However, one strength of this work is the demonstration of a direct correlation between the magnitudes of xylanase and glucanase activities throughout the entire GH5_4 subfamily (and potentially beyond), spanning a specific activity range of 4 orders of magnitude (see Figs. 3 and 5).

While a number of groups have reported on broad substrate specificity in GH5_4 and closely related enzymes [18–25], this work provides a more extensive, quantitative demonstration of a relationship between the two activities. For further validation of the relationships identified here, we examined data reported for 7 GH5 enzymes (mostly GH5_5) and 10 substrates in Vlasenko *et al.* [36]. In that work, activities on xylan correlated with activities on three types of cellulose (notably, however, the correlation with PASC was weak), in addition to corn stover, arabinoxylan, and xyloglucan. While several caveats apply to this secondary analysis, it is still notable that their correlations did not extend to either β-1,4-mannan or galactomannan.

The lack of a statistically supported connection with mannanase activity in this work adds intrigue to the xylanase–glucanase correlation. Mannan differs from cellulose (or the β-1,4 linked stretches of lichenan) in the positioning of the C2 hydroxyl, and it contrasts from xylan in retaining the C6 hydroxymethyl group. The axially orientated OH2 may impose geometric or energetic constraints on the −1 mannose along its path to the transition state, and as discussed more below, GH5_4 active sites may not be optimally structured to stabilize the mannose catalytic intermediate [11,37,38]. Structure–function studies with GH5 enzymes further suggest a preference for equatorial C2 hydroxyls at the −2 sugar binding site [39], and specificity may also arise from side chain interactions with OH3 in the −1 site, which, despite being similarly oriented for mannose and glucose in the ground state, adopts slightly different conformations during the catalytic itinerary [39,40]. Indeed, our observation of lower average temperature and pH optima for mannanases in GH5_4 might support a distinct binding or catalytic mechanism.

The pH and temperature optima might also hint at the significance of the axial C2 in mannan. The C2 hydroxyl for the glycosyl residue in the −1-site is located near two relatively well-conserved histidine residues [41]; in CelE, these are His148 and His149. In the case of residues comparable to His149, its replacement by Ala in a GH5_25 enzyme was shown to depress activity slightly more on galactomannan than on carboxymethyl cellulose (CMC) [26]. However, in another study, for an enzyme that is more closely related to GH5_4, replacement of either conserved His with Ala decreased activities by a comparable amount on both xylan and CMC [42], so the generality of the connection between His148, His149, and mannanase activity in particular is not clear. However, given that both lowered temperature and reduced pH thermodynamically drive His protonation, one can envision how protonation may facilitate a mechanistically favorable hydrogen bond with the C2 hydroxyl of mannan that cannot be achieved with lichenan.

There are also noteworthy differences in the substrates' macroscopic properties that may contribute to the specificity correlations. Notably, the high crystallinity of low DP β-1,4-mannan results in significantly reduced solubility [35], while lichenan and beechwood xylan both pack more loosely and form semi-soluble colloids in water. Many GH5

mannanases and cellulases are dependent upon auxiliary CBMs for interacting with crystalline substrates [43,44], but only the free enzyme domains were assayed here. As a result, only a small fraction of the mannan may be accessible, effectively lowering the apparent substrate concentration. Interestingly, the tight correlation between activities on galactomannan and an unbranched form of β-1,4-mannan reported in Vlasenko *et al*. (neither of which correlated with xylanase activity) suggests our choice of β-1,4-mannan instead of galactomannan as the representative substrate may not have mattered [36]. Altogether, some combination of mechanistic differences and material properties is the most likely reason for the (apparent) lack of correlation of glucanase or xylanase activity with mannanase activity.

In the context of this discussion, it seems noteworthy that in clades 1B and 2B, a statistically significant correlation is observed between lichenase and mannanase activity (but not xylanase; Table 2). Comparison of active site architectures in enzymes from these clades may reveal structural features linked to bifunctional glucanase–mannanase enzymes, providing contrast with the apparently more generally observed glucanase–xylanase bifunctionality. Further analysis of the sequences and active site structures in clade 3B may also be revealing, as the data suggest the presence of features that simultaneously improve performance on all three substrates.

Xyloglucanase activity has been previously noted in several GH5_4 members [27,29–34], and activity on xyloglucan is thought to be a hallmark property of GH5_4 [9]. Given the often wider binding clefts in clades 1 and 2, we hypothesized that these enzymes might be more compatible with such a branched substrate. While several otherwise weakly active enzymes did show stronger reactivity with xyloglucan, including six members from clades 1 and 2 with no other detectable activities, the relationship appears to be complex as activities and magnitudes were spread throughout all of GH5_4. CelE, which belongs to clade 3B, was previously noted for its high xyloglucanase activity on plant biomass using glycome profiling [27], and we measured its specific xyloglucanase activity at 8.2 U/mg in the present screen.

The broader trend of the highest activities being present in clade 3 appears to hold for xyloglucan as well, and a small positive correlation between xyloglucanase and lichenase activity exists (see Supplemental Data). It appears that the factors that make enzymes highly active on lichenan also apply to xyloglucan to some degree. The fact that clade 1 possesses a disproportionate number of xyloglucan-inactive enzymes similarly correlates with lichenase and xylanase behavior, although it has previously been noted that some [34] but not all [33] members appear to have a xyloglucan-competent active site cleft in the absence of the CBM46 module.

Outside of GH5_4, we observed xyloglucanase activity for both subfamily 25 members (GenBank accession codes ABS61403 and WP_004082283; see Fig. 5), while it was lacking in the more distantly related GH5_2 and GH5_5 representatives and minimal in the GH5_37 representative. Thus, we suggest that the appearance of xyloglucanase reactivity may have occurred slightly prior to the emergence of subfamily 4.

GH5_4 and domain modularity

Domain modularity provides another input controlling GH5 specificity. Several natural examples of GH5 proteins fused to GH26 domains have been reported, such as the GH5_25 cellulosomal *Ct*Lic26A-Cel5E [45] and the uncharacterized β-1,4-endoglucanase from *Prevotella* sp. Sc00026 (classified in GH5_4 but not assigned to any clade, i.e., “region I”). While GH26 is classically considered a mannanase family, *Ct*Lic26A has no activity on mannan, but the GH26–GH5 fusion displays two to three times the catalytic efficiency on lichenan than either domain alone. Another example encompassed within this work is PbGH5A from *Prevotella bryantii* B14, of which only the C-terminal GH5_4 domain has been crystallized (PDB ID 3VDH and others) and extensively characterized [32]. We found that this clade 3G enzyme has lichenase and xylanase activity but no mannanase activity, and the function of this protein's GH26 domain is unknown. Three other proteins in clade 3G share this general domain structure, and their lone GH5_4 domains also have no mannanase activity (Supplemental Data). A contrasting example is *Cb*Cel9B/Man5A, which is one of several proteins secreted from the thermophile *Caldicellulosiruptor bescii* featuring a GH5_1 domain that *confers* mannanase activity [46]. GH5 subfamilies 1, 2, 7, and 8 have been noted for having multi-domain fusions to CBMs and GH domains from other families, resulting in either mono- or bispecific enzymes, and dual cellulase-mannanases have been reported [9]. Therefore, while our work on the GH5_4 domains in isolation has revealed a compelling link between lichenase and xylanase activity, there is more to understand about how multiple domains cooperate to determine substrate specificity even in this single subfamily.

This study focused on GH5_4 catalytic cores instead of their native domain context in order to map functional measurements to the catalytic cores. GH5_4 is highly diverse in terms of the type and arrangement of domains of its member polypeptides, and the CBM46 (CBM X2)-containing endoglucanases of clade 1 were mentioned previously. We previously showed that fusion of clade 3 CelE from *H. thermocellum* to CBMs with different polysaccharide binding specificities could differentially enhance reactivity with lichenan, xylan, or mannan [27,44]. Several gene products in clades 2 and 3 employ the canonical linker spacing between the catalytic core and associated CBM modules, which

include cellulose-binding CBM2 and CBM10, chitin-binding CBM1, the mixed-affinity CBM4, and xyloglucan-specific CBM65. In all, 29 enzymes from clades 2 and 3 in our tested set were derived from constructs possessing CBMs. Notably, no consistent difference was observed between this pool of enzymes and those without native CBMs (see Supplemental Data).

GH5_4 optimization and specializations

The glycoside hydrolase family–subfamily organizational schema greatly simplifies the prediction and classification of new and existing enzymes by grouping them with evolutionarily related enzymes [8,9,47–49]. However, for large families or subfamilies, significant diversity in activity profiles is possible; for GH5_4, we have shown that maximum specific activities on lichenan and xylan span nearly 5 orders of magnitude. Further subcategorization within subfamilies can provide valuable clarity: here we observed that one of the three major clades in GH5_4 (clade 3) yields consistently higher activities on all polysaccharide substrates. Sub-subcategorization also revealed interesting finer features such as the above-noted possible evolution toward a bifunctional glucanase–mannanase enzyme (clades 1B and 2B), or the potential early stages of specialization toward either glucanase (clade 3A) or xylanase (clade 3G) activities.

Translating the phylogeny paradigm to applied scenarios such as the design of improved enzymes or selection of an optimal candidate for industrial use generally requires the challenging task of identifying only those clade-specific evolutionary changes tied to activity changes. Toward this aim, we examined sequence and structural features common to the best-performing groupings. In addition to the strong performance of enzymes in clade 3, four of the five enzymes from outside of GH5_4 also had high specific activities on all three unbranched substrates. At a subclade level, clade 2B was exceptional within clade 2.

To guide the discussion below, Fig. 6 presents the structure of *H. thermocellum* CelE (PDB 4IM4), with an emphasis on the scope of conservation of key substrate-binding residues. These residues are largely contributed by loop regions joining the C-terminus of a core β-strand with the subsequent α-helix. The most highly conserved of these residues are Asn192, Tyr270, and Trp349; these and the catalytic glutamates are among the seven residues that are seemingly entirely conserved throughout all of GH5 [50,51]. His148 is also highly conserved for a large swath of GH5, although the conservation does not extend to more distant subfamilies such as the mannanases in GH5_7 (e.g., Ref. [52]). Further narrowing in scope, Asn72 and Trp82 are observed universally only within GH5_4 and its most closely related families; the conservation of Trp82 extends furthest and includes GH5_2. The remaining positions (His149, Trp203, Tyr273, Asn351, and Glu360) display variability even within GH5_4.

Figure 7 summarizes the relationships among clade membership, enzymatic activity, and the identities of three notable binding site residues from that variable category. Along with CelE positions 149 and 203, the figure maps the residue corresponding to Tyr273 in CelE to each enzyme's location in the phylogenetic tree. The phylogenetic group-calculated cumulative enzyme activity distributions on the right-hand side of the figure emphasize the relatively strong performance by most clade 3 enzymes on lichenan, xylan, and mannan. As an example, enzymes in clade 3B are observed to meet or exceed the average enzyme in this study both in terms of fraction exceeding the detection threshold (leftmost strip in each subplot) and fraction exceeding any given activity value (main section of each subplot).

His149 was first observed to map to elevated mannanase activity by Chen *et al.* [26]. In Fig. 7, the solid purple bars indicate complete conservation of His149 throughout clades 1 and 3. In contrast, in clade 2, the residue identity is variable for all subclades besides 2B, which displayed higher activities than the other members of clade 2 in our assays on unbranched polysaccharides (see Fig. 7). In clades 2A and 2D, His149 is largely replaced by a tryptophan or tyrosine, while in clade 2C, it is replaced by glycine. The change in clade 2C is accompanied by a structural change in which loop 3 (joining strand 3 to canonical helix 3) is extended by approximately 4 amino acid residues. The structures of the enzyme from *Paenibacillus pabuli* (accession code AAR65335, PDB 2JEP and 2JEQ) illustrate this change [30]: rather than directly contacting the −1 glycosyl residue, the glycine in this position assists in formation of a short helical turn that extends outward from the core β-barrel, effectively increasing the contribution of loop 3 to the binding cleft. Two structures from clade 2D illustrate an alternative evolutionary route, in which the replacement residue is aromatic (PDBs 3ZMR and 4W8B [29,31]); in PDB 4W8B, the replacement tryptophan is observed to stack below the −1-branched xylosyl residue in the co-crystallized xyloglucan fragment. It is possible that the changes may be linked to changes in xyloglucan specificity or compatibility in clades 2C and 2D [29–31,53], and in fact, all enzymes in our limited xyloglucan test from these subclades did show activity on xyloglucan.

Equally interesting is the +1-stacking tryptophan residue in loop 4 (Trp203 in CelE, Fig. 6B). This position has been rationalized to facilitate binding of the substrate or to stabilize the enzyme–product complex [11]. Trp203 has been observed in all clade 2B and clade 3 structures to date, and Fig. 7 highlights its apparent conservation in these phylogenetic groups against the variability observed throughout the rest of GH5_4. In clade 1, loop 4 is generally shortened to such an extent that there is no homologous residue (see

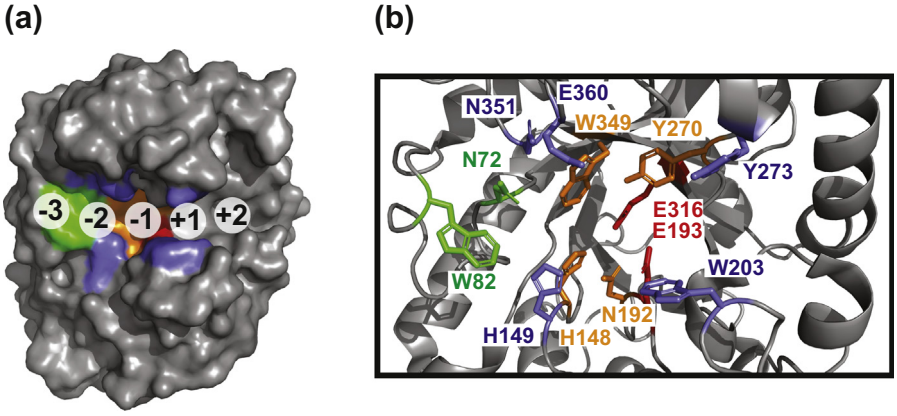

Fig. 6. Structure of CelE from *H. thermocellum* (PDB ID 4IM4). Residues previously implicated in substrate binding are highlighted according to the extent of conservation throughout GH5. Red, catalytic glutamates (E193 and E316); orange, GH5-wide conserved residues (H148, N192, Y270, W349); green, GH5_4 conserved residues (N72 and W82); slate, variable residue positions (H149, W203, Y273, N351, E360). (A) Surface representation of CelE showing the oligosaccharide binding channel. (B) Expanded view of the binding channel. Oligosaccharide subsite locations were identified through homology alignment with other GH5_4 crystal structures; in particular, 5D9N was used to identify subsites +1 and + 2, and an unpublished cellotriose-bound structure (CBL16772.1) was used for subsites −3, −2, and − 1.

PDB 4YZP [33]). The only exceptions are a small subset of enzymes in clade 1B that are most closely related to the ancestral clade 1 enzyme; perhaps coincidentally, both clade 1B crystal structures are from this region of phylogenetic space (Fig. 7, PDBs 5E09 and 4V2X [34]). In clade 2, the residue at 203 is heterogeneous, mirroring the scenario observed for His149. Trp203 is fully conserved in clade 2B (PDB 4YHE [54]), while in clades 2A and 2C, it is sometimes or usually replaced by a tyrosine, respectively. In clade 2D, an accompanying and likely related gross structural change further suggests evolution toward some other activity. Here loop 6, which lies opposite loop 4 in the binding cleft, is increased significantly in length relative to the rest of GH5_4, while loop 4 is contrastingly shortened. In enzymes closely related to the clade 2D ancestor, alignment suggests a tryptophan or tyrosine in this location, while the shortened loop 4 eliminates the structural position in the more distant members of 2D. Overall, Trp203 appears to play a compelling binding and/or catalytic role, and indeed, its conservative replacement by tyrosine in parts of clade 2 largely underscores the functional importance of this position. Another aromatic residue contributed by loop 6 on the other side of the channel (Tyr273 in CelE) has similarly been proposed to further stabilize substrate binding, perhaps acting in concert with the loop 4 aromatic residue [32,54]. In clade 3, this is often a tyrosine (as observed in CelE), while particularly strong conservation of tryptophan at this position is observed in clades 1 and 2C. Still, the clearest observation from this work is that conservation of Trp203 correlates with strong activity, suggesting that this residue comprises part of an optimally structured active site.

Representatives from related subfamilies in our data set (GH5_25, GH5_37, GH5_2, GH5_5; see Figs. 4 and 5B) lack both His149 and Trp203, and yet, four of the five enzymes possess catalytic activities and breadth of specificity similar to those in clade 3. His149 is preserved in subfamily 25 but varies for the others. In the case of Trp203, a tryptophan is instead provided by loop 5. This loop 5 tryptophan is highly conserved through GH5_2 and closely related families, and its structural/functional replacement by Trp203 in subfamily 4 has been noted for some time [50,55].

More work is needed to fully understand several other alterations affecting specificity changes, but there are some notable observations. Enzymes in the glucanase-specializing clade 3A possess an extended loop 4 with a second tryptophan, immediately following Trp203. In combination with the loop 6 tyrosine (Tyr273; see Fig. 7) on the opposite face of the binding cleft, this additional tryptophan creates a particularly extensive array of aromatic residues lining the channel's positive subsites (PDBs 1EDG [56], 4NF7, and 4X0V [21]). This may allow for a more specialized, substrate-specific binding channel in some enzymes. Indeed, while enzymes in clade 3A were responsible for several of this study's largest specific activity measurements on lichenan, they also comprise three of the four undetectable or marginally detectable xyloglucanase measurements in clade 3.

In contrast to clade 3A, the structural representative from the xylanase-specializing clade 3G (*P. bryantii* enzyme with accession code AAC97596.1) shows a binding cleft that may be more sterically compatible with branched substrates than the rest of clade 3 (PDB 3VDH) [32]. Loops 7 and 8 are respectively increased and decreased in length in clade 3G enzymes, and an Asp residue in loop 7 functionally replaces analogous residues in loop 8 that are otherwise conserved throughout clade 3 (Asn351 and Glu360 in CelE). One

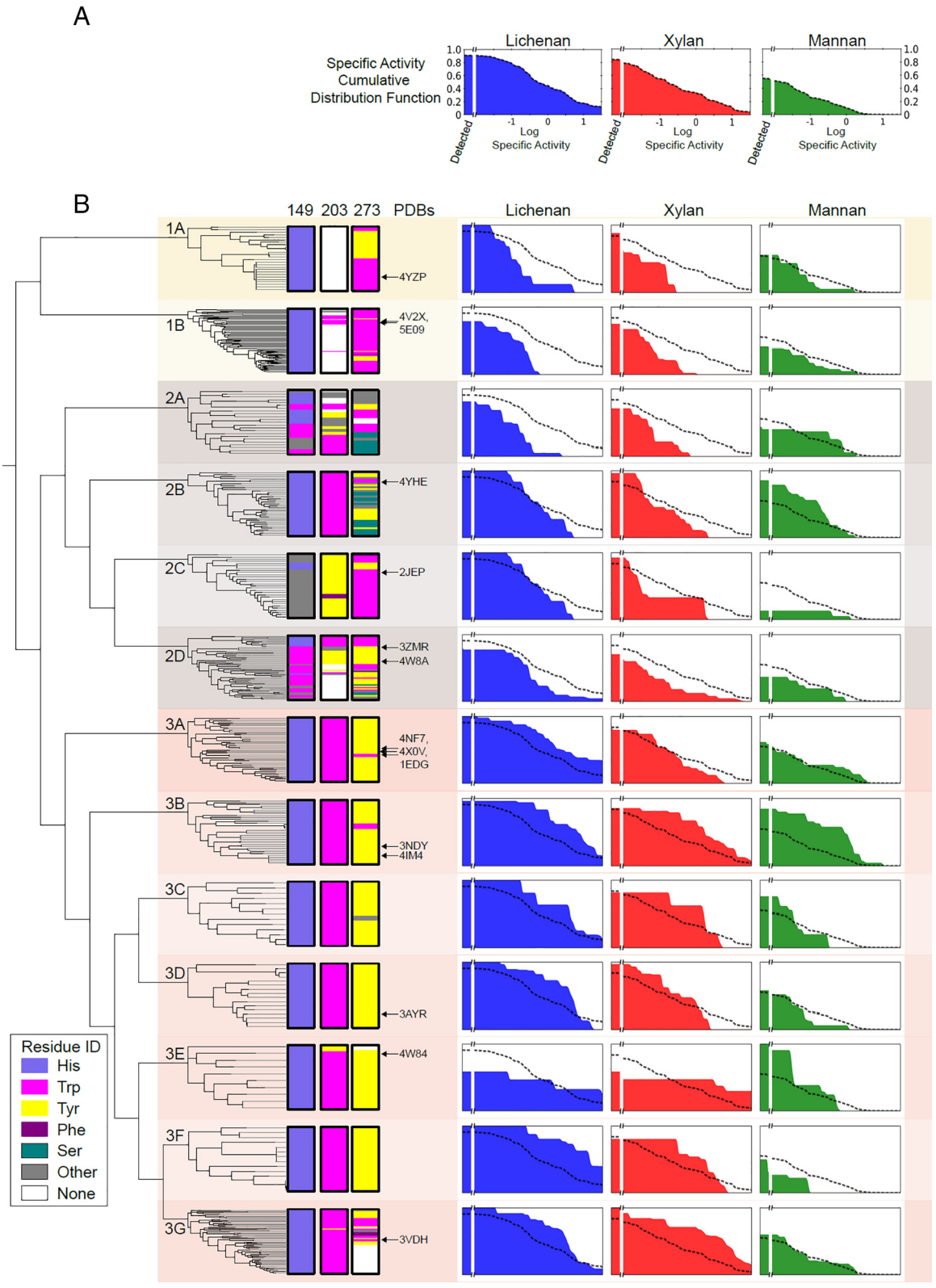

Fig. 7 (*legend on next page*)

possible result of this exchange is the creation of steric compatibility for oligosaccharides branching in the space formerly occupied by loop 8 [32]. Although we were able to align the simple xyloglucan fragments from the related PDB entries 5D9N and 5D9M into the binding channel from other structures in clade 3 such as CelE (4IM4) without noticeable clashing, and our assessment of xyloglucanase activity throughout GH5_4 indicates that most clade 3 enzymes have no trouble hydrolyzing xyloglucan, it is possible that the changes in clade 3G enable accommodation of additional complexity in xyloglucan branching (beyond a single α-1,6-linked xylosyl residue). Further contrasting with the rest of the clade, enzymes in 3G show significant variability at position 273. In AAC97596.1, a tryptophan serves as a functional replacement. In enzymes resulting from more recent branchings (bottom of Fig. 7), an analogous amino acid is lacking, likely due to significant shortening of loop 6. It is not clear yet whether these observations relate to the increased xylanase activity observed for this or the related subclades 3E and 3F.

In addition to possessing His149 and Trp203 homologs, clade 2B enzymes have a narrow and extended binding groove that is reminiscent of clade 3 enzymes [54], providing a structural distinction from the rest of clade 2 (one previous phylogenetic tree in fact identified clade 2B as distinct from the rest of clade 2 [33]). This is produced by the increase, relative to the rest of clade 2, in the average lengths for clade 2B enzyme loops proximal to the negative subsites (loops 1, 3, and 7). While clade 2B enzymes show lower specific activities than the top performers in clade 3 in the conditions of this study, their performance relative to the rest of clade 2 is noteworthy. In contrast, as for clade 3A, it is possible that the narrowed binding cleft increases the potential for incompatibility with branched substrates, as clade 2B contributes the only two xyloglucan-inactive representatives from clade 2. The lichenase–mannanase correlation in clade 2B may also indicate an interesting point of departure in function for enzymes in 2B relative to clade 3.

The evolution of GH5_4

Interestingly, the lengthening of loop 6 that occurs in clades 2C and particularly 2D described above likely represents a reversion to a more distant ancestral protein morphology. Comparison of alignments and structures in subfamilies 25, 36, 37, 38, 39, 52, and 2 suggests prominent loop 6 extensions as the rule. This contrasts with much of subfamily 4, where loop 4 is often increased in length, along with the aforementioned GH5_4 hallmark introduction of Trp203 in loop 4 [50,55].

Some understanding of the evolutionary path traversed from a more distant ancestral protein to the GH5_4 progenitor can be gained by examining sequences and structures closely related to ancestral enzymes along this lineage. CtCel5E (genbank accession ABN52701.1) is another *H. thermocellum* enzyme for which a structure has been solved (PDB 4U3A) [42], and it is annotated as subfamily 4 in the CAZy database. In the phylogenetic mapping here, this enzyme is outside of the three main clades and is descended from the preceding lineage that is labeled region I in Fig. 4. CtCel5E thus provides an interesting snapshot of the transition from more distant subfamilies to subfamily 4. In contrast to structures from GH5_2, GH5_37, and GH5_25, the enzyme possesses a significantly shortened loop 6. At the same time, it exhibits neither a loop 5 tryptophan nor its replacement in loop 4. In addition, a distinctive non-prolyl cis-peptide bond involving the GH5-ubiquitous tryptophan (W349 in CelE) apparently converts to trans in region I; this trans bond appears in CtCel5E and is present all GH5_4 structures. A GH5_4-conserved Gly–Gly motif, at the non-catalytic end of the barrel between α-helix 4 and β-strand 5, is not present in the CtCel5E structure but is evident in some proteins in this region.

Notably, many enzymes in region I possess CBMs or other auxiliary domains (Supplemental Data), so it is possible that loss of the tryptophan in loop 5 occurred simultaneously with addition of some compensatory or synergistic module. Furthermore, co-evolution of a multi-domain protein might result in an enzyme that is less optimized as a single catalytic domain. Based on mapping sequence alignment data to the evolutionary tree, it appears Trp203 was introduced late in region I. Thus, it may have been present in the common ancestor to GH5_4, but it only became fixed in clades 2B and 3. According to this reasoning, additional evolutionary changes were necessary to re-capitulate the functional importance of this convergently evolved tryptophan in GH5_4.

Fig. 7. Distribution and relationship of specific activities and identities of variable binding cleft residues. (A) Cumulative distribution plots for all tested enzymes; the plot contours describe the probability that an enzyme selected at random would exceed value provided on the abscissa. Left of breakpoint, probability that an activity would be detected in our assays; right of breakpoint, probability that the activity would exceed a specific value. Blue, red, and green graphs represent measurements on lichenan, xylan, and mannan. (B) Schematic of binding channel composition and activity data for each clade. The higher-order phylogenetic tree branches are presented in topological format for clarity (relative tree branch lengths *within* each subclade are meaningful). Extension of a leaf node through the three bar plots reveals the alignment-determined identities of key binding channel residues corresponding to positions 149, 203, and 273 in CelE. The PDB column notes the location of enzymes with deposited crystallographic data. Cumulative-specific activity distribution functions are plotted on the right-hand side of each row; for reference, the contour for all tested enzymes is included as a dashed line in each subplot.

Conclusions

This subfamily-wide screen provides a novel view of the evolution and landscape of specificity in a glycoside hydrolase family. Most strikingly, this study revealed the synchronous tuning of glucanase and xylanase functions over an evolutionary timescale that generated divergent configurations of the substrate binding channel. Combined with structural and biochemical work from others, we also identified key residues associated with elevated bifunctional glucanase/xylanase activity, as well as an indication of more recent events mapping to the potential genesis of new functionalities and/or specificity profiles.

The observed correlations in activity suggest that many evolutionary alterations to an enzyme simultaneously alter the kinetics for multiple activities. Whether (and for which substrates) these synchronous changes typically occur in the substrate association term (i.e., K_M) or turnover number remains as a relevant question for understanding and improving enzyme function. Detailed comparative studies of representative enzymes that deconstruct kinetic parameters [34] or provide detailed product analyses [21,22,27,32] are required. Developing an understanding of the types of activities that are complementary *versus* oppositional will provide a new approach to the understanding and design of substrate specificity in glycoside hydrolases.

Materials and Methods

Gene synthesis and cell-free translation

GH5 catalytic domain sequence candidates were identified through BLAST search for sequences similar to CelE, with the goal of obtaining an enzyme pool representative of the sequence diversity in GH5_4. Optimal lengths were determined through iterative sequence alignment and comparison to crystal structures. Genes were synthesized at the Joint Genome Institute (Walnut Creek, CA) and cloned into the cell-free translation vector pEU (12). Enzymes were expressed by sequential cell-free transcription-translation using wheat germ extract from CellFree Sciences (Yokohama, Japan) as described in Takasuka *et al.* [28]. Briefly, DNA sequences were transcribed to mRNA using SP6 polymerase, and mRNAs were mixed with wheat germ extract and translated in a bilayered, diffusion-fed translation reaction. Product enzyme concentrations were measured by band analysis on stain-free SDS PAGE.

Enzymatic reactions in screen

Icelandic moss lichenan, tamarind seed xyloglucan, and borohydride-reduced β-(1,4)-mannan from Megazyme (Wicklow, Ireland), and beechwood xylan from Sigma (St. Louis, MO, USA) were prepared at 20 mg/mL in 0.05% sodium azide. Reactions (40 μL total volume) consisted of 4 μL of cell-free translation mixture containing enzyme, 16 μL 0.1 M phosphate or acetate buffer, and 20 μL of stock substrate. Plates were incubated on a PCR thermocycler without shaking for 2 h or overnight (~16 h). Each reaction was conducted at pH 6.0 at three different temperatures: 30, 50, and 70 °C. Reactions of lichenan and xylan were also conducted at 50 °C, at pH values 4, 5, 7, and 8. For reactions with mannan, the pH-varied assays were instead incubated at 30 °C due to the generally lower temperature optimum for mannanase activity. Xyloglucanase activity was screened in similar fashion as above, except that reactions were run for 3 h at 30 °C and pH 6.

To measure reducing sugars released from enzymatic hydrolysis reactions, 30 μL reaction supernatant was mixed with 60 μL DNS reagent (5.3 g 3,5-dinitrosalicylic acid, 9.9 g sodium hydroxide, 153 g sodium potassium tartrate, 4.2 g sodium metabisulfite, 3.8 mL phenol per liter aqueous solution) and heated at 95 °C for 5 min. Samples were read at 540 nm on a Tecan Infinite M1000 microplate reader. Threshold signals for activity assignments were established by performing enzyme-free controls ($n \geq 20$) for each set of pH, temperature, and substrate conditions; the minimal specific activity required for detection varied for each enzyme–substrate pair according to these conditions and enzyme expression levels. On average, the enzyme concentration produced through cell-free expression was 0.07 μg/μL, so the average microplate assays included 0.3 μg of protein. Thus, the average lower limit detectable specific activity was $\sim 5 \times 10^{-2}$ U/mg. At the upper end of the dynamic range, the maximum detectable activity in this screen is around 7×10^{-3} U (40% hydrolysis in 2 h), so that the average maximum specific activity is 25 U/mg.

Phylogram construction and sequence analysis

The phylogram was generated from MrBayes 3.2.6 [57]. Tree building employed a total of 638 sequences (sequences in addition to the tested 243 were extracted from the CAZy GH5 database according to their alignment to a customized Hidden Markov model) using two independent runs and achieved a final average standard deviation of split frequencies between runs of 0.0104. Nucleotide model settings were nst = 6 and rates = invgamma, and Metropolis-coupled MCMC utilized 64 chains, with Nswaps = 4, Swapfreq = 1, and the temperature parameter was generally set to 0.01. The tree was constructed using the MrBayes MPI implementation on UWCHTC HPC cluster. Gene domain architectures referred to in the main text and plotted in Supplemental Data were generated using HMMER (http://hmmer.org/) and profiles from the Pfam database [58–60]. For clarity in

the presentation of activity information, the tree in Fig. 4 removed sequences >65% in sequence identity to a neighboring node, unless it was part of the tested set.

A custom Hidden Markov Model was constructed for sequence alignment and comparison. The starting model was based heavily on clade 3 enzyme features. Clade 3 enzymes were aligned to this model first, and probabilities were updated by Baum–Welch. The resulting model was then used to align clade 2 sequences, and then similarly for clade 1. Changes in loop lengths referred to in Discussion were determined by setting a reference point in the alignment that included all loop densities.

Numerical analysis

To minimize systematic underestimation of activities for enzymes nearing the practical maximum yield for a substrate, activity values were back-extrapolated to theoretical values at 5% yield according to an estimated yield *versus* time curve (yield(t)). For this curve, we used a previously determined functional form for CelE on xylan, which is approximately described as an exponential decay with a maximum yield of 20% (i.e., a total reducing end population equivalent to 2 mg/mL glucose). For lichenan, the same functional form was applied but the maximum yield set to an empirically determined upper limit of 36%. In all cases, the time required to achieve 5% yield was estimated through data fitting yield(t). Upper and lower confidence measures were determined by back-extrapolating a 5% uncertainty in the measured yield value, and the corresponding bounds are presented in the scatterplots in Figs. 3 and 5. There is some necessary oversimplification in this analysis, but the intent is to avoid significant systematic underestimates in presentation. Only measurements at the upper end of lichenase and xylanase measurements are significantly impacted (see Supplemental Data). Only statistical trends that were unaffected by the correction are reported, including rank ordering of mean and median activities for all substrates, outliers in the lichenase–xylanase trend (3A and 3G), and statistically significant correlations presented in Table 2.

Averages, regression parameters, and confidence intervals (Tables 1 and 2) were determined by bootstrap sampling (the statistics for ≥1000 resulting resampled data sets are reported) as described below. Each specific activity was represented as a normal probability distribution with a breadth σ set to the uncertainty in the activity correction described above; activities that were likely lower limits were represented as broad, highly skewed distributions with $\sigma = 1$ log unit. Each sampling step consisted of selecting values from these probability distributions, nested within a bootstrap sampling step. Probability density plots in Fig. 7 were calculated using a similar sampling approach, except that the selection pool included all enzymes (even those with no detectable activities). All specific activity averaging and statistical analyses were performed with log scaling. Figures were generated with either Matplotlib [61] or Bokeh [62], and basic statistical tests were performed using the Scipy library [63].

The following accession numbers are referred to in this text: GenBank ID: AAR65335, GenBank ID: AAC97596, GenBank ID: ABN52701, GenBank ID: ABS61403, GenBank ID: WP_004082283, PDB ID: 4YZP, PDB ID: 4V2X, PDB ID: 5E09, PDB ID: 4YHE, PDB ID: 2JEP, PDB ID: 3ZMR, PDB ID: 4W8A, PDB ID: 4NF7, PDB ID: 4X0V, PDB ID: 1EDG, PDB ID: 3NDY, PDB ID: 4IM4, PDB ID: 3AYR, PDB ID: 4W84, PDB ID: 3VDH, PDB ID: 4W8B, PDB ID: 5D9N, PDB ID: 5D9M, and PDB ID: 4U3A.

Acknowledgments

The authors acknowledge the Department of Energy, Office of Science, for funding: DE-SC0018409 and DE-FC02-07ER64494 (GLBRC) and DE-AC02-05CH11231 (JGI). E.G. is supported by the NIGMS Biotechnology Training Program (NIH 5 T32 GM008349) at the University of Wisconsin–Madison.

This research was performed using the compute resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

Conflict of Interest: The authors declare no competing financial interests.

Author Contributions: T.T., E.G., K.V.M., and B.F. conceived the experiments; S.D. performed gene synthesis; E.G. performed the high-throughput screen; L.B. assisted in sample preparation; and K.V.M. performed the bioinformatics analyses. E.G., K.V.M., T.T., and B.F. wrote the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.01.024.

Received 10 September 2018;
Received in revised form 19 October 2018;
Accepted 16 January 2019
Available online 25 January 2019

Keywords:
glycoside hydrolase;
substrate specificity;
polysaccharide;
protein evolution;
synthetic biology

†E.M.G. and K.A.V.M. contributed equally to this work.

Abbreviations used:
GH5, glycoside hydrolase family 5; CAZy, Carbohydrate-Active enZyme; GH5_4, GH5 subfamily 4; CMC, carboxymethyl cellulose.

References

[1] O. Khersonsky, D.S. Tawfik, Enzyme promiscuity: a mechanistic and evolutionary perspective, Annu. Rev. Biochem. 79 (2010) 471–505.
[2] R.A. Jensen, Enzyme recruitment in evolution of new function, Annu. Rev. Microbiol. 30 (1976) 409–425.
[3] B. Hocker, J. Claren, R. Sterner, Mimicking enzyme evolution by generating new (beta alpha)(8)-barrels from (beta alpha)(4)-half-barrels, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 16448–16453.
[4] D.W. Banner, A.C. Bloomer, G.A. Petsko, D.C. Phillips, C.I. Pogson, I.A. Wilson, et al., Structure of chicken muscle triose phosphate isomerase determined Crystallographically at 2.5 Å resolution using amino-acid sequence data, Nature 255 (1975) 609–614.
[5] A.D. Goldman, J.T. Beatty, L.F. Landweber, The TIM barrel architecture facilitated the early evolution of protein-mediated metabolism, J. Mol. Evol. 82 (2016) 17–26.
[6] N. Nagano, C.A. Orengo, J.M. Thornton, One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions, J. Mol. Biol. 321 (2002) 741–765.
[7] M.V. Omelchenko, M.Y. Galperin, Y.I. Wolf, E.V. Koonin, Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution, Biol. Direct 5 (2010).
[8] V. Lombard, H.G. Ramulu, E. Drula, P.M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Res. 42 (2014) D490–D495.
[9] H. Aspeborg, P.M. Coutinho, Y. Wang, H. Brumer III, B. Henrissat, Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5), BMC Evol. Biol. 12 (2012) 186.
[10] S.G. Withers, Mechanisms of glycosyl transferases and hydrolases, Carbohydr. Polym. 44 (2001) 325–337.
[11] G.J. Davies, L. Mackenzie, A. Varrot, M. Dauter, A.M. Brzozowski, M. Schulein, et al., Snapshots along an enzymatic reaction coordinate: analysis of a retaining beta-glycoside hydrolase, Biochemistry 37 (1998) 11707–11713.
[12] S.G. Burton, Oxidizing enzymes as biocatalysts, Trends Biotechnol. 21 (2003) 543–549.
[13] B. Joseph, P.W. Ramteke, G. Thomas, Cold active microbial lipases: some hot issues and recent developments, Biotechnol. Adv. 26 (2008) 457–470.
[14] T. Schafer, T.W. Borchert, V.S. Nielsen, P. Skagerlind, K. Gibson, K. Wenger, et al., Industrial enzymes, Adv. Biochem. Eng. Biotechnol. 105 (2007) 59–131.
[15] X.W. Peng, H. Su, S.F. Mi, Y.J. Han, A multifunctional thermophilic glycoside hydrolase from *Caldicellulosiruptor owensensis* with potential applications in production of biofuels and biochemicals, Biotechnol. Biofuels 9 (2016).
[16] N. Srivastava, M. Srivastava, P.K. Mishra, V.K. Gupta, G. Molina, S. Rodriguez-Couto, et al., Applications of fungal cellulases in biofuel production: advances and limitations, Renew. Sust. Energ. Rev. 82 (2018) 2379–2386.
[17] N. Annamalia, M.V. Rajeswari, T. Balasubramanian, Endo-1,4-β-Glucanases: Role, Applications and Recent Developments, 2016.
[18] C.M. Bianchetti, P. Brumm, R.W. Smith, K. Dyer, G.L. Hura, T.J. Rutkoski, et al., Structure, dynamics, and specificity of endoglucanase D from *Clostridium cellulovorans*, J. Mol. Biol. 425 (2013) 4267–4285.
[19] H.P. Fierobe, C. Gaudin, A. Belaich, M. Loutfi, E. Faure, C. Bagnara, et al., Characterization of endoglucanase a from *Clostridium cellulolyticum*, J. Bacteriol. 173 (1991) 7956–7962.
[20] J. Liu, C. Tsai, J. Liu, K. Cheng, C. Cheng, The catalytic domain of a *Piromyces rhizinflata* cellulase expressed in *Escherichia coli* was stabilized by the linker peptide of the enzyme, Enzym. Microb. Technol. 28 (2001) 582–589.
[21] D.D. Meng, X. Liu, S. Dong, Y.F. Wang, X.Q. Ma, H. Zhou, et al., Structural insights into the substrate specificity of a glycoside hydrolase family 5 lichenase from *Caldicellulosiruptor* sp. F32, Biochem. J. 474 (2017) 3373–3389.
[22] M. Iakiviak, R.I. Mackie, I.K. Cann, Functional analyses of multiple lichenin-degrading enzymes from the rumen bacterium *Ruminococcus albus* 8, Appl. Environ. Microbiol. 77 (2011) 7541–7550.
[23] F.C.F. Foong, R.H. Doi, Characterization and comparison of *Clostridium cellulovorans* endoglucanases–xylanases Engb and Engd hyperexpressed in *Escherichia coli*, J. Bacteriol. 174 (1992) 1403–1409.
[24] E. Berger, W.A. Jones, D.T. Jones, D.R. Woods, Cloning and sequencing of an endoglucanase (End1) gene from *Butyrivibrio fibrisolvens* H17c, Mol. Gen. Genet. 219 (1989) 193–198.
[25] G.P. Xue, K.S. Gobius, C.G. Orpin, A novel polysaccharide hydrolase Cdna (Celd) from *Neocallimastix patriciarum* encoding 3 multifunctional catalytic domains with high endoglucanase, cellobiohydrolase and xylanase activities, J. Gen. Microbiol. 138 (1992) 2397–2403.
[26] Z. Chen, G.D. Friedland, J.H. Pereira, S.A. Reveco, R. Chan, J.I. Park, et al., Tracing determinants of dual substrate specificity in glycoside hydrolase family 5, J. Biol. Chem. 287 (2012) 25335–25343.
[27] J.A. Walker, S. Pattathil, L.F. Bergeman, E.T. Beebe, K. Deng, M. Mirzai, et al., Determination of glycoside hydrolase specificities during hydrolysis of plant cell walls using glycome profiling, Biotechnol. Biofuels 10 (2017).
[28] T.E. Takasuka, J.A. Walker, L.F. Bergeman, K.A. Vander Meulen, S. Makino, N.L. Elsen, et al., Cell-free translation of biofuel enzymes, Methods Mol. Biol. 1118 (2014) 71–95.
[29] C.R. Dos Santos, R.L. Cordeiro, D.W. Wong, M.T. Murakami, Structural basis for xyloglucan specificity and alpha-D-Xylp(1 → 6)-D-Glcp recognition at the −1 subsite within the GH5 family, Biochemistry 54 (2015) 1930–1942.
[30] T.M. Gloster, F.M. Ibatullin, K. Macauley, J.M. Eklof, S. Roberts, J.P. Turkenburg, et al., Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12, J. Biol. Chem. 282 (2007) 19177–19189.
[31] J. Larsbrink, T.E. Rogers, G.R. Hemsworth, L.S. McKee, A.S. Tauzin, O. Spadiut, et al., A discrete genetic locus confers

xyloglucan metabolism in select human gut Bacteroidetes, Nature 506 (2014) 498–502.
[32] N. McGregor, M. Morar, T.H. Fenger, P. Stogios, N. Lenfant, V. Yin, et al., Structure–function analysis of a mixed-linkage beta-glucanase/xyloglucanase from the key ruminal Bacteroidetes *Prevotella bryantii* B(1)4, J. Biol. Chem. 291 (2016) 1175–1197.
[33] M.V. Liberato, R.L. Silveira, E.T. Prates, E.A. de Araujo, V.O. Pellegrini, C.M. Camilo, et al., Molecular characterization of a family 5 glycoside hydrolase suggests an induced-fit enzymatic mechanism, Sci. Rep. 6 (2016), 23473.
[34] I. Venditto, S. Najmudin, A.S. Luis, L.M. Ferreira, K. Sakka, J.P. Knox, et al., Family 46 carbohydrate-binding modules contribute to the enzymatic hydrolysis of xyloglucan and beta-1,3-1,4-glucans through distinct mechanisms, J. Biol. Chem. 290 (2015) 10572–10586.
[35] L.R. Moreira, E.X. Filho, An overview of mannan structure and mannan-degrading enzyme systems, Appl. Microbiol. Biotechnol. 79 (2008) 165–178.
[36] E. Vlasenko, M. Schulein, J. Cherry, F. Xu, Substrate specificity of family 5, 6, 7, 9, 12, and 45 endoglucanases, Bioresour. Technol. 101 (2010) 2405–2411.
[37] G. Speciale, A.J. Thompson, G.J. Davies, S.J. Williams, Dissecting conformational contributions to glycosidase catalysis and inhibition, Curr. Opin. Struct. Biol. 28 (2014) 1–13.
[38] D.J. Vocadlo, G.J. Davies, Mechanistic insights into glycosidase chemistry, Curr. Opin. Chem. Biol. 12 (2008) 539–555.
[39] L.E. Tailford, V.M.A. Ducros, J.E. Flint, S.M. Roberts, C. Morland, D.L. Zechel, et al., Understanding how diverse beta-mannanases recognize heterogeneous substrates, Biochemistry 48 (2009) 7009–7018.
[40] L.E. Tailford, W.A. Offen, N.L. Smith, C. Dumon, C. Morland, J. Gratien, et al., Structural and biochemical evidence for a boat-like transition state in β-mannosidases, Nat. Chem. Biol. 4 (2008) 306.
[41] T.H. Wu, C.H. Huang, T.P. Ko, H.L. Lai, Y. Ma, C.C. Chen, et al., Diverse substrate recognition mechanism revealed by *Thermotoga maritima* Cel5A structures in complex with cellotetraose, cellobiose and mannotriose, Biochim. Biophys. Acta 1814 (2011) 1832–1840.
[42] S.F. Yuan, T.H. Wu, H.L. Lee, H.Y. Hsieh, W.L. Lin, B. Yang, et al., Biochemical characterization and structural analysis of a bifunctional cellulase/xylanase from *Clostridium thermocellum*, J. Biol. Chem. 290 (2015) 5739–5748.
[43] D. Hogg, G. Pell, P. Dupree, F. Goubet, S.M. Martin-Orue, S. Armand, et al., The modular architecture of *Cellvibrio japonicus* mannanases in glycoside hydrolase families 5 and 26 points to differences in their role in mannan degradation, Biochem. J. 371 (2003) 1027–1043.
[44] J.A. Walker, T.E. Takasuka, K. Deng, C.M. Bianchetti, H.S. Udell, B.M. Prom, et al., Multifunctional cellulase catalysis targeted by fusion to different carbohydrate-binding modules, Biotechnol. Biofuels 8 (2015).
[45] E.J. Taylor, A. Goyal, C.I.P.D. Guerreiro, J.A.M. Prates, V.A. Money, N. Ferry, et al., How family 26 glycoside hydrolases orchestrate catalysis on different polysaccharides—structure and activity of a *Clostridium thermocellum* lichenase, CtLic26A, J. Biol. Chem. 280 (2005) 32761–32767.
[46] R. Wang, L. Gong, X.L. Xue, X. Qin, R. Ma, H.Y. Luo, et al., Identification of the C-terminal GH5 domain from CbCel9B/Man5A as the first glycoside hydrolase with thermal activation property from a multimodular bifunctional enzyme, PLoS One (2016) 11.
[47] K. Mewis, N. Lenfant, V. Lombard, B. Henrissat, Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization, Appl. Environ. Microbiol. 82 (2016) 1686–1692.
[48] F.J. St John, J.M. Gonzalez, E. Pozharski, Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups, FEBS Lett. 584 (2010) 4435–4441.
[49] M.R. Stam, E.G. Danchin, C. Rancurel, P.M. Coutinho, B. Henrissat, Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins, Protein Eng. Des. Sel. 19 (2006) 555–562.
[50] J. Sakon, W.S. Adney, M.E. Himmel, S.R. Thomas, P.A. Karplus, Crystal structure of thermostable family 5 endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose, Biochemistry 35 (1996) 10648–10660.
[51] Q. Wang, D. Tull, A. Meinke, N.R. Gilkes, R.A. Warren, R. Aebersold, et al., Glu280 is the nucleophile in the active site of *Clostridium thermocellum* CelC, a family a endo-beta-1,4-glucanase, J. Biol. Chem. 268 (1993) 14096–14102.
[52] A.M. Goncalves, C.S. Silva, T.I. Madeira, R. Coelho, D. de Sanctis, M.V. San Romao, et al., Endo-beta-D-1,4-mannanase from *Chrysonilia sitophila* displays a novel loop arrangement for substrate selectivity, Acta Crystallogr. D Biol. Crystallogr. 68 (2012) 1468–1478.
[53] M.A. Attia, C.E. Nelson, W.A. Offen, N. Jain, G.J. Davies, J.G. Gardner, et al., In vitro and in vivo characterization of three *Cellvibrio japonicus* glycoside hydrolase family 5 members reveals potent xyloglucan backbone-cleaving functions, Biotechnol. Biofuels 11 (2018) 45.
[54] A.E. Naas, A.K. MacKenzie, B. Dalhus, V.G. Eijsink, P.B. Pope, Structural features of a Bacteroidetes-affiliated cellulase linked with a polysaccharide utilization locus, Sci. Rep. 5 (2015), 11666.
[55] G.J. Davies, M. Dauter, A.M. Brzozowski, M.E. Bjornvad, K.V. Andersen, M. Schulein, Structure of the *Bacillus agaradherans* family 5 endoglucanase at 1.6 angstrom and its cellobiose complex at 2.0 angstrom resolution, Biochemistry 37 (1998) 1926–1932.
[56] V. Ducros, M. Czjzek, A. Belaich, C. Gaudin, H.P. Fierobe, L.P. Belaich, et al., Crystal-structure of the catalytic domain of a bacterial cellulase belonging to family-5, Structure 3 (1995) 939–949.
[57] F. Ronquist, M. Teslenko, P. van der Mark, D.L. Ayres, A. Darling, S. Hohna, et al., MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space, Syst. Biol. 61 (2012) 539–542.
[58] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[59] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37.
[60] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, et al., The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res. 44 (2016) D279–D285.
[61] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95.
[62] Team BD, Bokeh: Python Library for Interactive Visualization, 2018.
[63] E. Jones, T. Oliphant, P. Peterson, et al., Scipy: Open Source Scientific Tools for Python, 2001.