Supplemental Data

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

Authors:

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]

Affiliations:

[1]Great Lakes Bioenergy Research Center, Madison, WI, USA

[2]Department of Biochemistry, University of Wisconsin – Madison, USA

[3]Research Faculty of Agriculture, Hokkaido University, Sapporo, Japan

[4]Department of Chemistry, University of Wisconsin – Oshkosh, USA

[5]DOE Joint Genome Institute, Walnut Creek, CA, USA

A Distinct Activity Distribution for Clade 3 Enzymes — Supplementary Figure S1 illustrates that activity values in Clade 3 enzymes are shifted toward higher activities for lichenan and xylan. For mannanase activities, the assays are likely sampling from the upper end of a distribution and so limit definitive conclusions.

Compared to the other substrates, activities on xyloglucan show an overall increase in average activities. The histogram is also more compressed than those of lichenan and xylan, possibly due to the fact that fully soluble xyloglucan is more easily degraded than the other substrates, so enzymes with activity on xyloglucan are approaching yield saturation. Most relevant to this work, the activities indicate a rank-ordering (Clade 3 > Clade 2 > Clade 1) in median activity that matches the results for lichenan and xylan. Thus it appears that the general catalytic efficiency exhibited by an enzyme in assays of lichenase or xylanase activities transfers to the hydrolysis of xyloglucan. The possibility that open binding clefts in Clades 1 and 2 make these enzymes more generalists with regard to branched substrates remains open, as the tamarind xyloglucan used in this study represents only one possible xyloglucan structure.

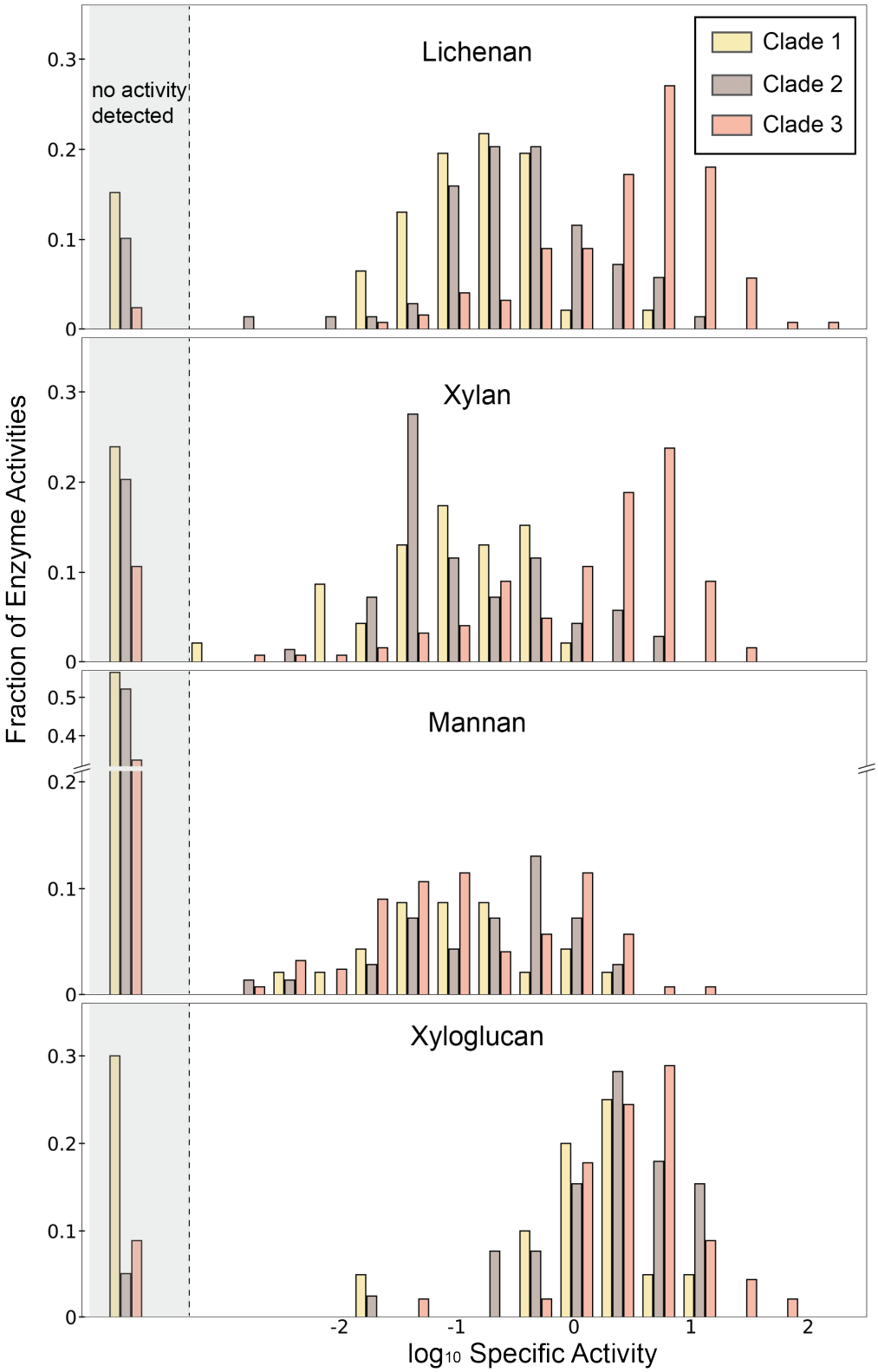

Figure S1. Per-clade histograms of specific activities on lichenan, xylan, mannan, or xyloglucan. Because the xyloglucan data are not as extensive, activity data for all substrates is presented not corrected for substrate saturation. Histograms are presented in terms of fraction of the tested population.

Other correlations — Table S1 reports correlations observed following activity measurements in a second set of assays collected with additional substrates (tamarind xyloglucan, carboxymethylcellulose (CMC), phosphoric acid swollen cellulose (PASC) and beechwood xylan). All reactions in this smaller dataset were carried out for 3 hours at 30 °C, pH 6.

Table S1. Correlation Statistics for Additional Substrates and Lichenan or Xylan

		Lichenan [a]				Xylan [a]			
Substrate [b]	Enz.[c]	Pts.[d]	p [e]	ρ [f]	slope [f]	Pts.[d]	p [e]	ρ [f]	slope [f]
Xyloglucan	110	90	$< 10^{-3}$	0.35 (0.08 – 0.42)	0.0 (0.0 – 0.5)	82	0.7	0.04 (-0.16 – 0.19)	0.0 (-0.3 – 0.3)
CMC	64	47	$< 10^{-6}$	0.65 (0.51 – 0.76)	0.6 (0.5 – 0.9)	43	0.01	0.38 (0.12 – 0.61)	0.4 (0.0 – 0.9)
PASC	46	20	0.07	0.41 (0.04 – 0.69)	0.5 (0.1 – 1.1)	20	0.6	-0.11 (-0.48 – 0.28)	0.3 (-0.4 – 1.0)
Xylan	64	60	$< 10^{-5}$	0.56 (0.38 – 0.71)	0.8 (0.6 – 1.0)	53	$< 10^{-4}$	0.52 (0.30 – 0.70)	0.8 (0.6 – 1.1)

a: Specific activities measured on lichenan and xylan in the main text complete assay (243 enzymes)
b: Substrate assayed in second experimental subset
c: Number of enzymes assayed in second experimental subset
d: Number of datapoints in the log-log comparison dataset (enzymes with both activities > 0)
e: p-value, Spearman ρ, regression slope
f: Best estimate and 90% confidence interval

The strongest correlations observed are between specific activities measured on carboxymethylcellulose and lichenan, and between xylan and either lichenan or xylan (the latter serves as an internal validation, since protein translations and substrate preparations were prepared entirely independently of the first set of experiments). Correlations between activities on xyloglucan or PASC and lichenan are also detectable, with the former characterized as rather shallow (log-log slope ≤ 0.5).

In the full dataset, we do not detect a significant subfamily-wide correlation between activities on mannan and either lichenan or xylan. If there is such a correlation, as mentioned above, both the average activities and histograms suggest it is stronger with lichenan (Figure S2).

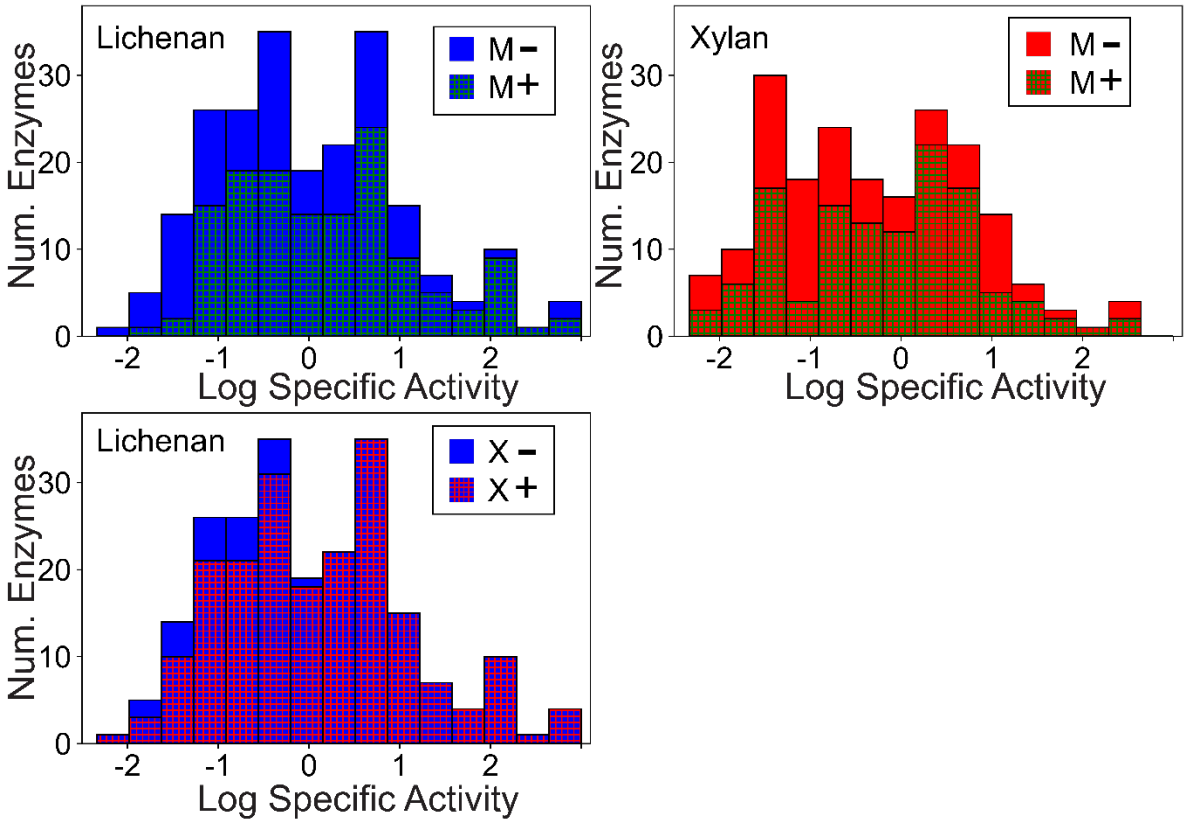

Figure S2. Histograms of specific activities on lichenan (A and C) or xylan (B), stacked to display frequencies for enzymes with or without mannanase activity (M+ vs M-, panels (A) and (B)) or frequencies for enzymes with or without xylanase activity (X+ vs X-, panel (B)).

Impact of CBM modules — No obvious functional difference is observed between enzymes possessing or lacking native CBM modules in Clades 2 and 3. Histograms in Figure S3 were generated by merging data from all sub-clades in 2 and 3 following subtraction of the corresponding median activity for that sub-clade. No systematic shift is observable, and a significant difference was not detected (via t-test) between the groups for either xylanase or lichenase activities. While it is possible there are sub-clades or examples in this population where enzyme function is significantly hampered by examining only the catalytic core, the data does not provide strong evidence for this. Moreover, none of the enzymes with no detectable activity were from gene constructs possessing CBM modules (see below).

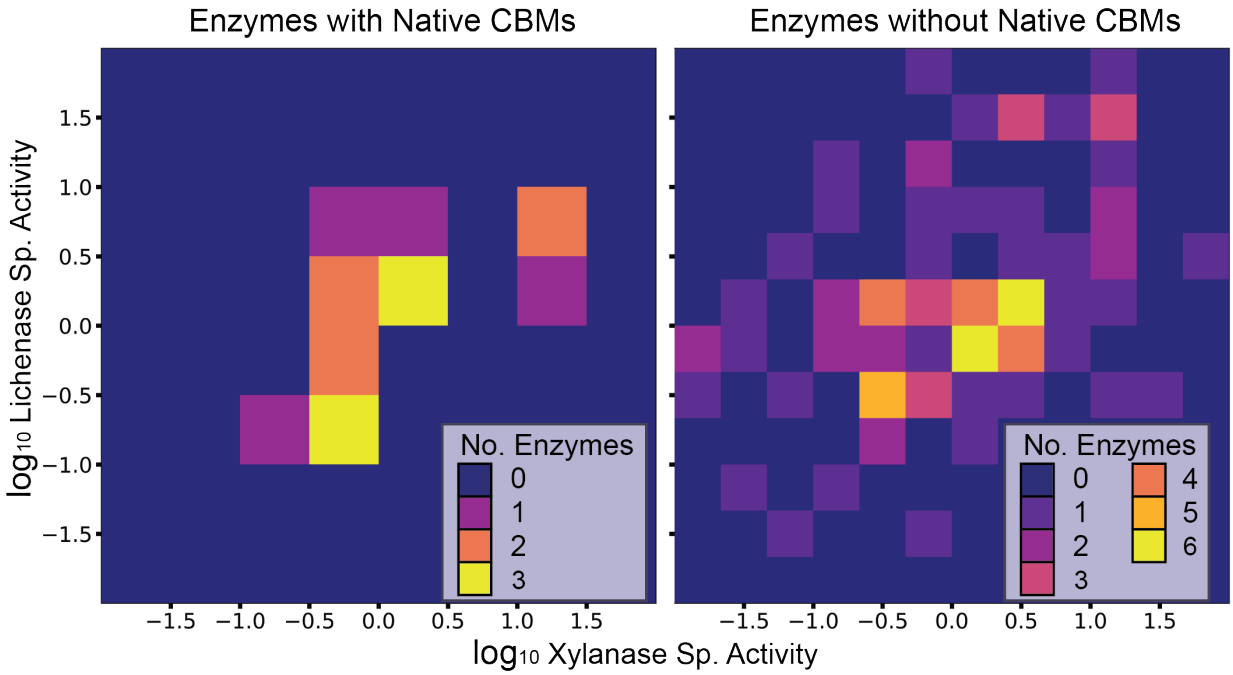

Figure S3. 2D histograms plotting the number of Clade 2 and 3 enzymes in a given span of lichenase and xylanase activity, relative to the median activity in the phylogenetic group for each enzyme. For enzymes originating from genes coding for CBMs in addition to the catalytic core (left), bin sizes are 0.5 log units; for enzymes originating from genes without CBMs, bin sizes are 0.33 log units.

In addition to CBMs, genes coding for the enzymes in this study may possess many other types of domains, such as protein localization bacterial immunoglobulin ("Big") or bacteroidetes-associated carbohydrate-binding often N-terminal (BACON) domains, while other GH5_4 members harbor dockerin domains for cellulosome association. As mentioned in the main text, several GH5_4 catalytic domains

are fused to other enzymatic domains. These can be other GH5_4 domains (up to four total), a GH26 domain, or a carbohydrate esterase domain, and the functions of these auxiliary enzyme domains may be related plant biomass degradation, such as simultaneous hydrolysis of cellulose and the tightly associated hemicelluloses, or deacetylation of polysaccharide side chains. As was the case with GH5_4 at the outset of this work, many of these protein families remain mostly under-characterized, and the functions assigned to them are typically based on one or a few examples. A systematic examination of GH5_4 with the full suite of additional domains observed in nature is beyond the scope of this work but would likely yield tremendous insight into the proteins' true function and utility.

Compensation for Substrate Depletion and Comparison to Previous Data — One nuance for the lichenase – xylanase correlation is that some measurements (10% of xylanase data, 22% of lichenase data) were performed after substrates depletions of 10% or more, meaning those data are outside of the linear region typically employed in enzyme kinetics studies (Iakiviak, et al. 2011; McGregor, et al. 2016). Beyond this 10% cutoff, the steady-state approximation becomes invalid and the reaction velocity decreases, so that a single-point specific activity measurement would underestimate the enzyme activity relative to determination in the steady-state regime. As described in Methods, we therefore adjusted enzymes in the main text to a specific activity corresponding to the hypothetical experimental time point corresponding to 5% substrate depletion. This adjustment only significantly impacts data collected after depletions of 10% or more (Figure S4).

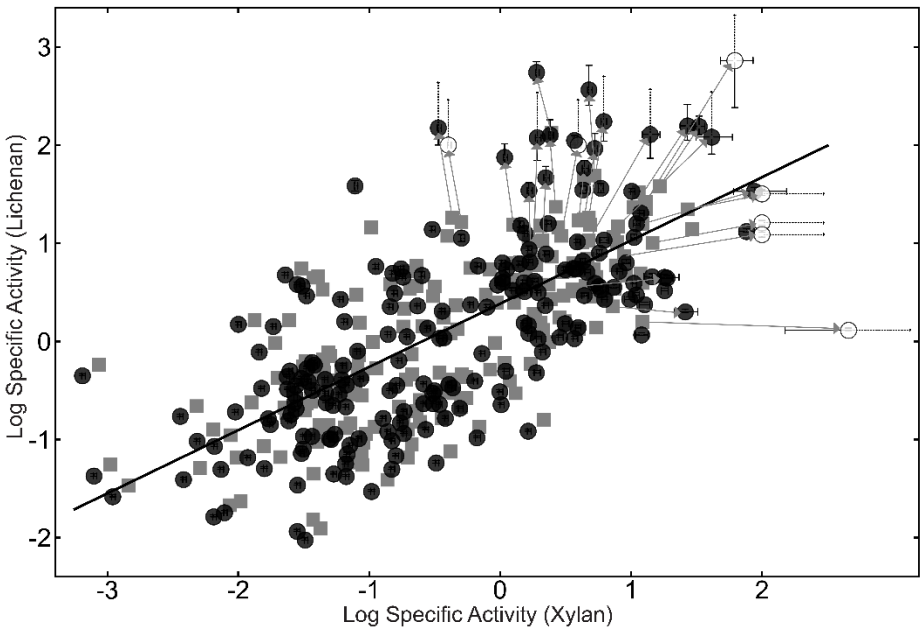

Figure S4. Lichenase vs xylanase correlation plot pre- and post-adjustment to hypothetical value at 5% substrate depletion. Gray squares, simple specific activity calculations; closed and open circles, adjusted data as plotted in main text Figure 4. Arrows displayed if the plot location is altered by more than 0.5 log units.

Encouragingly, our results also show a consistent relationship to literature measurements throughout the entire dynamic range. Figure S5 displays pre- and post-adjusted data in comparison with previous literature measurements. Measured specific activities on lichenan and xylan are roughly 10 – 100-fold lower in our assays than in several previous studies (Fierobe, et al. 1991; Foong and Doi 1992; Xue, et al. 1992; Liu, et al. 2001; Bianchetti, et al. 2013; McGregor, et al. 2016; Meng, et al. 2017). This is attributed to the lack of shaking in the high-throughput plate assays used here, differences due to scale of reaction, and possibly uncertainties in protein concentration calculations.

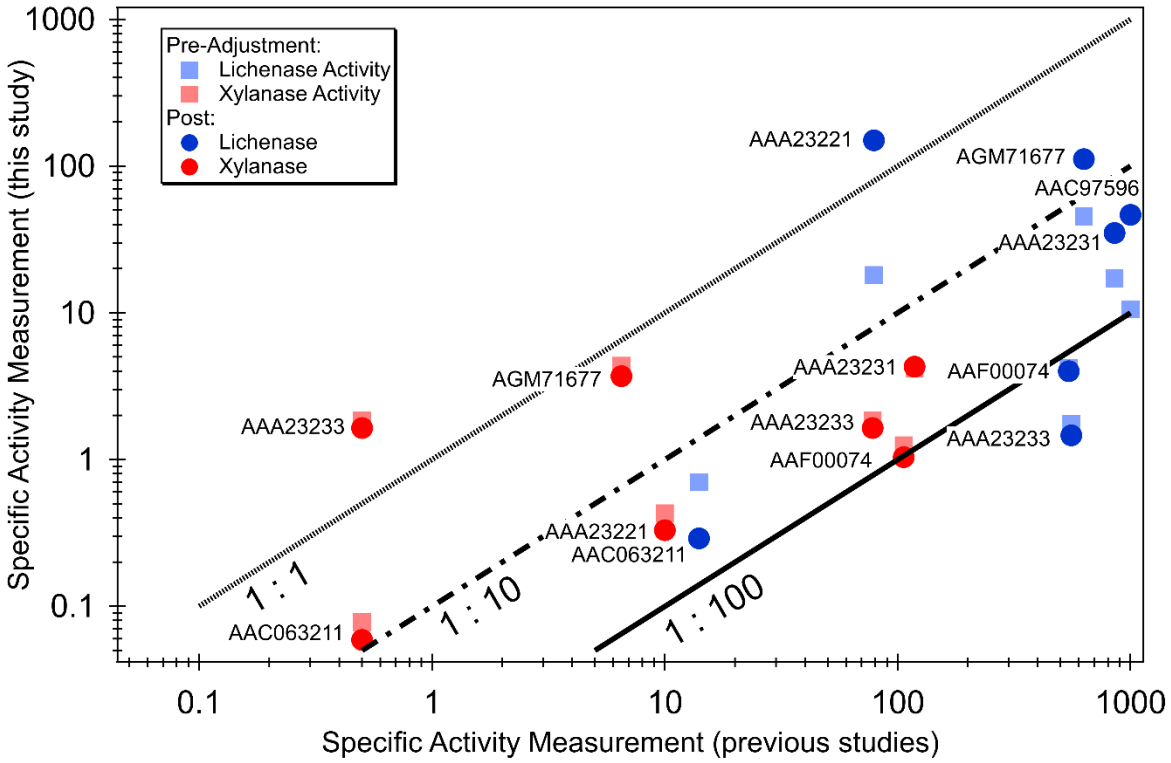

Figure S5. Comparison between specific activity values in quantitative screen and measurements in previous studies

Complete phylograms for each main group with detailed annotation are displayed below in Figure S6A – D (figures corresponding to related subfamilies, and Clades 1, 2, and 3). Bayesian MCMC probabilities ('supports') are listed above each node. Activity annotations from this study are included to the right of each node tip as in Figure 4 of the main text, with an added purple square similarly denoting that xyloglucanase activity was measured. As for the other activities, the size of the xyloglucanase symbol corresponds to its activity lying within the first, second or third quantile. EC numbers listed in the CAZy database are also listed. In the informational table at the right-hand side of the figure, the accession code and Pfam motif annotation for each gene are listed. Motifs are plotted to highlight features most relevant to this work: purple rectangles, GH5; yellow rectangles, other GH families; purple triangles, CBM X2 motif (a.k.a. CBM46 and the related Ig-like domain); tan triangles, other CBM motifs; brown ellipses, all other motifs. Locations of crystal structures and representative PDB accession codes are also marked between the relevant node tip and table entry.

Letters in the final column list isolation notes as provided in NCBI database: W= aquatic sample, S=terrestrial sample; R= rumen or other host-associated sample. Aside from handful of examples, Clade 1 enzymes have been found in soil or aquatic samples. Clades 2 and 3 are heavily represented by rumen and gut symbionts; of evolutionary note, even within subclades 2C, 2D and 3B, which are dominated by free living organisms, the enzymes that are most closely related to the common ancestor come from host-associated organisms.

Closely-related subfamilies and study outgroup representatives from subfamily 2 (AAC19169.1) and subfamily 5 (AFY52522.1) were included in the experimental set and the tree-building process. In Region I, there is a representative crystal structure corresponding to the protein with Genbank accession code ABN52701.1. This sequence was not included in the tree-building process, but it is most closely related to ACL75216.1 and AKP16689.1. It is marked as an added hypothetical node – table connecting line.

Fig. S6A

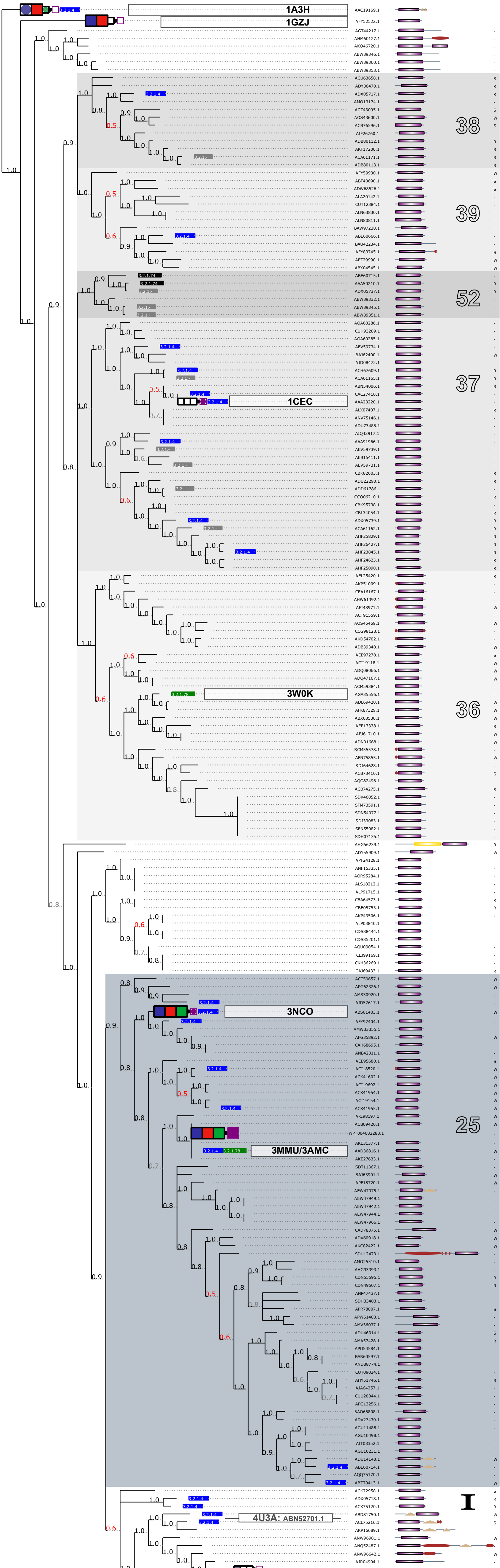
1A3H
1GZJ
1CEC
3W0K
3NCO
3MMU/3AMC
4U3A: ABN52701.1
38
39
52
37
36
25
I

Fig. S6B

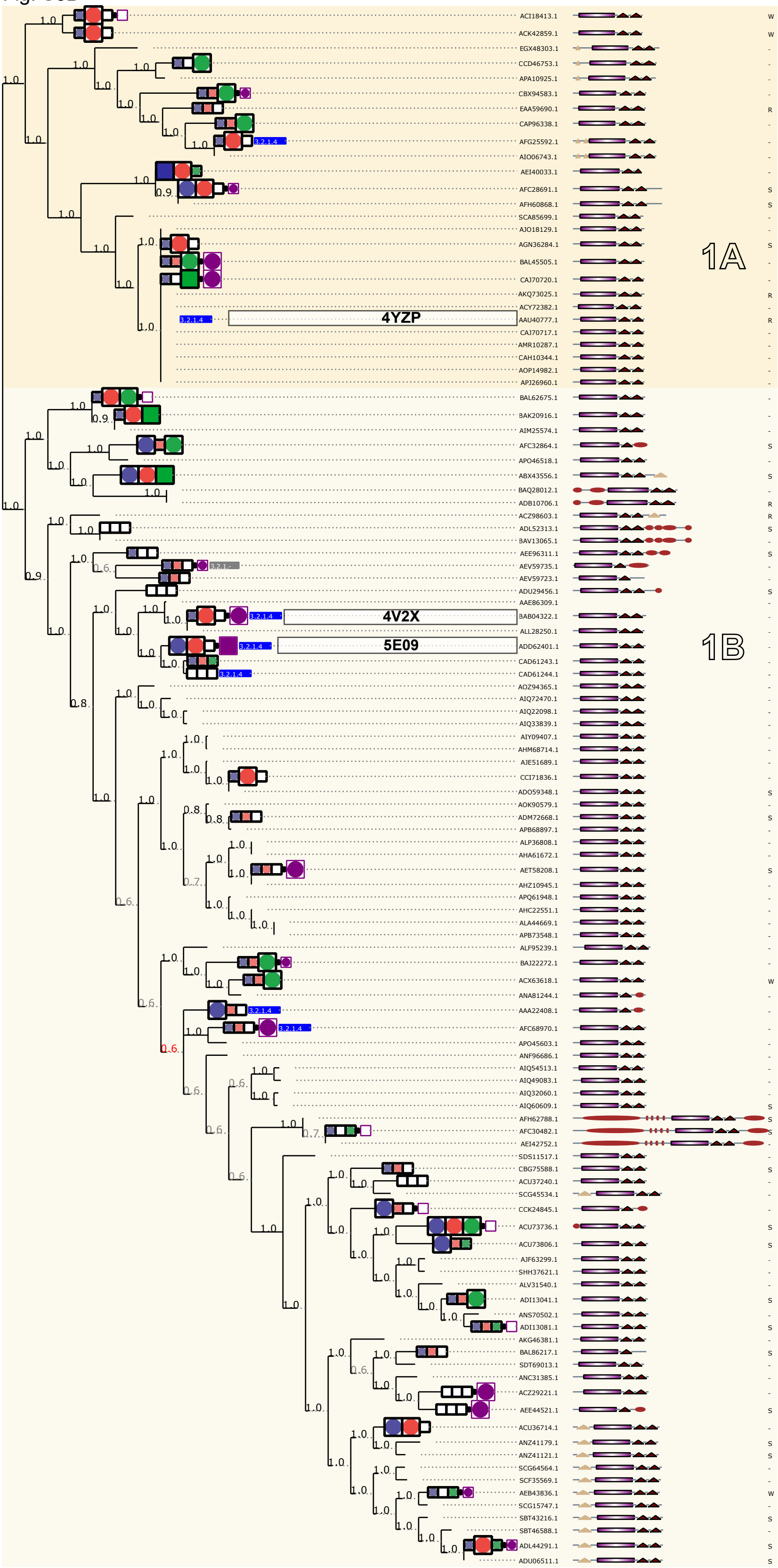
4YZP
4V2X
5E09
1A
1B
ACI18413.1
ACK42859.1
EGX48303.1
CCD46753.1
APA10925.1
CBX94583.1
EAA59690.1
CAP96338.1
AFG25592.1
AIO06743.1
AEI40033.1
AFC28691.1
AFH60868.1
SCA85699.1
AJO18129.1
AGN36284.1
BAL45505.1
CAJ70720.1
AKQ73025.1
ACY72382.1
AAU40777.1
CAJ70717.1
AMR10287.1
CAH10344.1
AOP14982.1
APJ26960.1
BAL62675.1
BAK20916.1
AIM25574.1
AFC32864.1
APO46518.1
ABX43556.1
BAQ28012.1
ADB10706.1
ACZ98603.1
ADL52313.1
BAV13065.1
AEE96311.1
AEV59735.1
AEV59723.1
ADU29456.1
AAE86309.1
BAB04322.1
ALL28250.1
ADD62401.1
CAD61243.1
CAD61244.1
AOZ94365.1
AIQ72470.1
AIQ22098.1
AIQ33839.1
AIY09407.1
AHM68714.1
AJE51689.1
CCI71836.1
ADO59348.1
AOK90579.1
ADM72668.1
APB68897.1
ALP36808.1
AHA61672.1
AET58208.1
AHZ10945.1
APQ61948.1
AHC22551.1
ALA44669.1
APB73548.1
ALF95239.1
BAJ22272.1
ACX63618.1
ANA81244.1
AAA22408.1
AFC68970.1
APO45603.1
ANF96686.1
AIQ54513.1
AIQ49083.1
AIQ32060.1
AIQ60609.1
AFH62788.1
AFC30482.1
AEI42752.1
SDS11517.1
CBG75588.1
ACU37240.1
SCG45534.1
CCK24845.1
ACU73736.1
ACU73806.1
AJF63299.1
SHH37621.1
ALV31540.1
ADI13041.1
ANS70502.1
ADI13081.1
AKG46381.1
BAL86217.1
SDT69013.1
ANC31385.1
ACZ29221.1
AEE44521.1
ACU36714.1
ANZ41179.1
ANZ41121.1
SCG64564.1
SCF35569.1
AEB43836.1
SCG15747.1
SBT43216.1
SBT46588.1
ADL44291.1
ADU06511.1

Fig. S6C

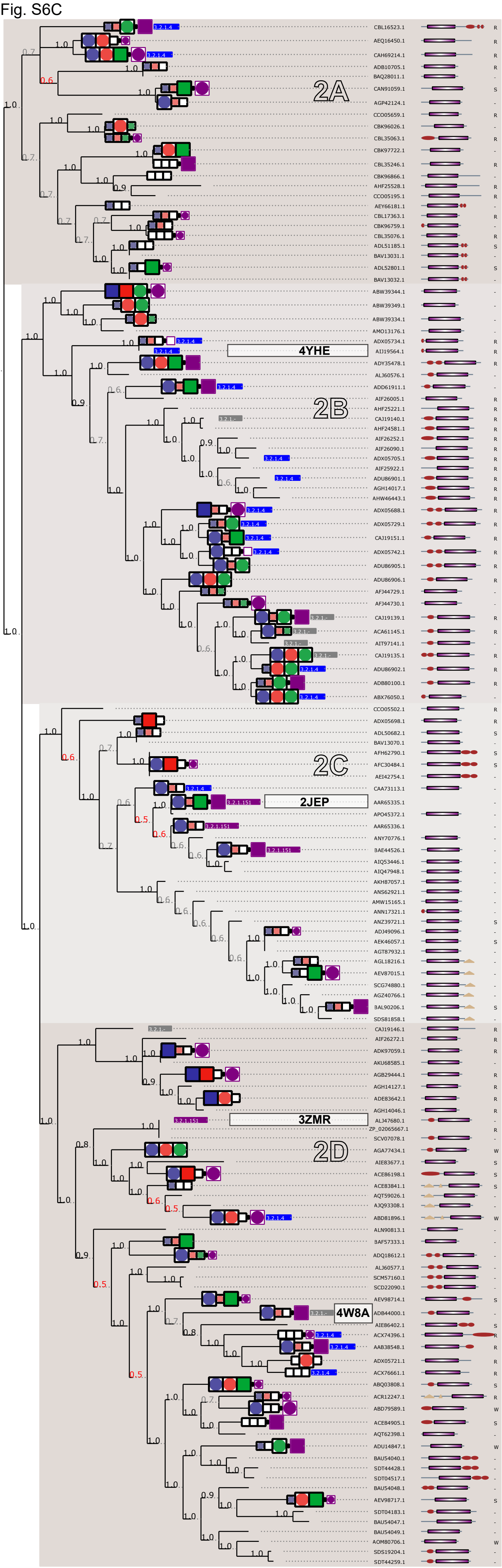
2A
2B
2C
2D
4YHE
2JEP
3ZMR
4W8A
CBL16523.1
AEQ16450.1
CAH69214.1
ADB10705.1
BAQ28011.1
CAN91059.1
AGP42124.1
CCO05659.1
CBK96026.1
CBL35063.1
CBK97722.1
CBL35246.1
CBK96866.1
AHF25528.1
CCO05195.1
AEY66181.1
CBL17363.1
CBK96759.1
CBL35076.1
ADL51185.1
BAV13031.1
ADL52801.1
BAV13032.1
ABW39344.1
ABW39349.1
ABW39334.1
AMO13176.1
ADX05734.1
AIJ19564.1
ADY35478.1
ALJ60576.1
ADD61911.1
AIF26005.1
AHF25221.1
CAJ19140.1
AHF24581.1
AIF26252.1
AIF26090.1
ADX05705.1
AIF25922.1
ADU86901.1
AGH14017.1
AHW46443.1
ADX05688.1
ADX05729.1
CAJ19151.1
ADX05742.1
ADU86905.1
ADU86906.1
AFJ44729.1
AFJ44730.1
CAJ19139.1
ACA61145.1
AIT97141.1
CAJ19135.1
ADU86902.1
ADB80100.1
ABX76050.1
CCO05502.1
ADX05698.1
ADL50682.1
BAV13070.1
AFH62790.1
AFC30484.1
AEI42754.1
CAA73113.1
AAR65335.1
APO45372.1
AAR65336.1
ANY70776.1
BAE44526.1
AIQ53446.1
AIQ47948.1
AKH87057.1
ANS62921.1
AMW15165.1
ANN17321.1
ANZ39721.1
ADJ49096.1
AEK46057.1
AGT87932.1
AGL18216.1
AEV87015.1
SCG74880.1
AGZ40766.1
BAL90206.1
SDS81858.1
CAJ19146.1
AIF26272.1
ADK97059.1
AKU68585.1
AGB29444.1
AGH14127.1
ADE83642.1
AGH14046.1
ALJ47680.1
ZP_02065667.1
SCV07078.1
AGA77434.1
AIE83677.1
ACE86198.1
ACE83841.1
AQT59026.1
AJQ93308.1
ABD81896.1
ALN90813.1
BAF57333.1
ADQ18612.1
ALJ60577.1
SCM57160.1
SCD22090.1
AEV98714.1
ADB44000.1
AIE86402.1
ACX74396.1
AAB38548.1
ADX05721.1
ACX76661.1
ABQ03808.1
ACR12247.1
ABD79589.1
ACE84905.1
AQT62398.1
ADU14847.1
BAU54040.1
SDT44428.1
SDT04517.1
BAU54048.1
AEV98717.1
SDT04183.1
BAU54047.1
BAU54049.1
AOM80706.1
SDS19204.1
SDT44259.1
3.2.1.4
3.2.1.151
3.2.1.-

Fig. S6D

References

Bianchetti CM, Brumm P, Smith RW, Dyer K, Hura GL, Rutkoski TJ, Phillips GN, Jr. 2013. Structure, dynamics, and specificity of endoglucanase D from Clostridium cellulovorans. J Mol Biol 425:4267-4285.
Fierobe HP, Gaudin C, Belaich A, Loutfi M, Faure E, Bagnara C, Baty D, Belaich JP. 1991. Characterization of endoglucanase A from Clostridium cellulolyticum. J Bacteriol 173:7956-7962.
Foong FCF, Doi RH. 1992. Characterization and Comparison of Clostridium-Cellulovorans Endoglucanases-Xylanases Engb and Engd Hyperexpressed in Escherichia-Coli. Journal of Bacteriology 174:1403-1409.
Iakiviak M, Mackie RI, Cann IK. 2011. Functional analyses of multiple lichenin-degrading enzymes from the rumen bacterium Ruminococcus albus 8. Appl Environ Microbiol 77:7541-7550.
Liu J, Tsai C, Liu J, Cheng K, Cheng C. 2001. The catalytic domain of a Piromyces rhizinflata cellulase expressed in Escherichia coli was stabilized by the linker peptide of the enzyme. Enzyme Microb Technol 28:582-589.
McGregor N, Morar M, Fenger TH, Stogios P, Lenfant N, Yin V, Xu X, Evdokimova E, Cui H, Henrissat B, et al. 2016. Structure-Function Analysis of a Mixed-linkage beta-Glucanase/Xyloglucanase from the Key Ruminal Bacteroidetes Prevotella bryantii B(1)4. J Biol Chem 291:1175-1197.
Meng DD, Liu X, Dong S, Wang YF, Ma XQ, Zhou H, Wang X, Yao LS, Feng Y, Li FL. 2017. Structural insights into the substrate specificity of a glycoside hydrolase family 5 lichenase from Caldicellulosiruptor sp. F32. Biochem J 474:3373-3389.
Xue GP, Gobius KS, Orpin CG. 1992. A Novel Polysaccharide Hydrolase Cdna (Celd) from Neocallimastix-Patriciarum Encoding 3 Multifunctional Catalytic Domains with High Endoglucanase, Cellobiohydrolase and Xylanase Activities. Journal of General Microbiology 138:2397-2403.