

# Structural and functional characterization of a ruminal β-glycosidase defines a novel subfamily of glycosyl hydrolase family 3 with permuted domain topology

Ramirez-Escudero, Mercedes; del Pozo, Mercedes V.; Marin-Navarro, Julia; Gonzalez, Beatriz; Golyshin, Peter; Polaina, Julia; Ferrer, Manuel; Sanz-Aparico, Julia

**Journal of Biological Chemistry**

*DOI:*
10.1074/jbc.M116.747527

Published: 11/11/2016

Peer reviewed version

**Manuscript Title:** Structural and Functional Characterization of a Ruminal β-glycosidase Defines a Novel Subfamily of Glycosyl Hydrolase Family 3 with Permuted Domain Topology

**Manuscript No:** JBC/2016/747527 [R2]

**Manuscript Type:** Regular Paper

**Date Submitted by the Author:** 16 Sep 2016

**Complete List of Authors:** Mercedes Ramirez-Escudero, Mercedes V. Del Pozo, Julia Marin-Navarro, Beatriz Gonzalez, Peter N. Golyshin, Julio Polaina, Manuel Ferrer, and Julia Sanz-Aparicio

**Keywords:** enzyme structure; glycoside hydrolase; metagenomics; protein structure ; X-ray crystallography; GH3 family; b-glycosidase; enzymology ; permuted domains topology; phylogenetic analysis

# Structural and functional characterization of a ruminal β-glycosidase defines a novel subfamily of glycoside hydrolase family 3 with permuted domain topology

**Mercedes Ramírez-Escudero[a], Mercedes V. del Pozo[b], Julia Marín-Navarro[c], Beatriz González[a], Peter N. Golyshin[d,e], Julio Polaina[c], Manuel Ferrer[b#] and Julia Sanz-Aparicio[a#]**

[a]*Department of Crystallography and Structural Biology, Inst. of Physical-Chemistry "Rocasolano", CSIC, Serrano 119, 28006 Madrid, Spain*

[b]*Inst. of Catalysis and Petrochemistry, CSIC, Marie Curie 2, Cantoblanco, 28049 Madrid, Spain.*

[c]*Inst. of Agrochemistry and Food Technology, CSIC, Carrer Catedràtic Agustín Escardino Benlloch 7, 46980 Paterna, Valencia, Spain*

[d]*School of Biological Sciences, Bangor University, LL57 2UW Gwynedd, UK*

[e]*Immanuel Kant Baltic Federal University, 236040, Kaliningrad, Russia*

Running title: Structure and function of a ruminal β−glycosidase

# To whom correspondence should be addressed. J. Sanz-Aparicio, Tel: (+34) 91 561 9400; e-mail: xjulia@iqfr.csic.es. M. Ferrer, Tel: (+34) 91 585 4872; e-mail: mferrer@icp.csic.es

**Keywords**: β-glycosidase; glycoside hydrolase family GH3; metagenomics; permuted domains topology, X-ray crystallography.

## ABSTRACT

Metagenomics has opened up a vast pool of genes for putative, yet uncharacterized, enzymes. It widens our knowledge on the enzyme diversity world and discloses new families for which a clear classification is still needed, as is exemplified by glycoside hydrolase (GH) family-3 proteins. Herein, we describe a GH3 enzyme ($GlyA_1$) from resident microbial communities in strained ruminal fluid. The enzyme is a β-glucosidase/β-xylosidase that also shows β-galactosidase, β-fucosidase, α-arabinofuranosidase and α-arabinopyranosidase activities. Short cello- and xylo-oligosaccharides, sophorose and gentibiose are among the preferred substrates, the large polysaccharide lichenan being also hydrolysed by $GlyA_1$. The determination of the crystal structure of the enzyme in combination with deletion and site-directed mutagenesis allowed identifying its unusual domain composition and the active site architecture. Complexes of $GlyA_1$ with glucose, galactose and xylose allowed picturing the catalytic pocket and illustrated the molecular basis of the substrate specificity. A hydrophobic platform defined by residues Trp711 and Trp106, located in a highly mobile loop, appears able to allocate differently β-linked bioses. $GlyA_1$ includes an additional C-terminal domain previously unobserved in GH3 members, but crystallization of the full-length enzyme was unsuccessful. Therefore, small angle x-ray experiments have been performed to investigate the molecular flexibility and overall putative shape. This study provided evidences that $GlyA_1$ defines a new subfamily of GH3 proteins with a novel permuted domain topology. Phylogenetic analysis indicates that this topology is associated with microbes inhabiting the digestive tracts of ruminants and other animals, feeding on chemically diverse plant polymeric materials.

Family 3 of glycoside hydrolases (GH3) contains about 11,000 entries among which diverse enzyme activities including β-glucosidase, β-xylosidase, exo-chitosanase, β-N-acetylglucosaminidase, glucocerebrosidase, exo-1,4-β-glucosidase and exo-1,3/1,4-β-glucanase have been characterized (http://www.cazy.org (1)). Few reported cases are bifunctional α-L-arabinopyranosidase/β-galactosidase (2), N-acetyl-β-glucosaminidase/β-glucosidase (3), β-glucosidase/cellodextrinase (4), β-xylosidase/α-L-arabinofuranosidase (5) and β-glucosidase/β-xylosidase (6). They are retaining enzymes that remove single glycosyl residues from the non-reducing end of their substrates. Therefore, they perform catalysis by a two-steps mechanism through a covalent enzyme-glycon intermediate, which is subsequently hydrolysed via an oxocarbenium-ion-like transition state.

Despite the high number of known GH3 sequences, structural knowledge on members of the GH3 family was absent until 1999, when the three-dimensional structure of the β-D-glucan exohydrolase Exo1 from *Hordeum vulgare* (barley) was reported (7). This study showed the core structure of most GH3 enzymes consisting of an N-terminal $(\alpha/\beta)_8$ barrel domain 1, which houses the active site pocket and the nucleophile, and a C-terminal $(\alpha/\beta)_6$-sandwich domain 2, containing the acid/base catalyst. The contribution of different domains in supplying crucial catalytic residues was a highly unusual feature of GH3 enzymes. Furthermore, in the last few years many new structural studies have shown a great variety in domain composition and arrangement of typical GH3 β-glycosidases, having up to four separate domains (8-15). Although this variety produces a shift in the sequence position of the acid/base catalyst, the known structures revealed that its structural location is well conserved among the different members. In contrast, several reported structures have revealed a more uniform pattern of the β-N-acetylglucosaminidases (NagZ) members showing that, despite a few having two-domains, most Gram-negative bacteria encode single domain enzymes, and all of them have the acid/base catalyst in an unusual histidine/aspartate dyad located in a flexible loop of the $(\alpha/\beta)_8$ barrel (16). This highly mobile loop has been proved to participate in substrate distortion to a $^1S_3$ conformation, therefore forming a productive Michaelis complex along catalysis (17). This has not been observed in other GH3 enzymes, with the substrate being in a relaxed chair conformation, although a Michaelis complex has been recently reported for the *Listeria innocua* β-glucosidase (18). Among all GH3 β-glycosidases with available structures, insights into the substrate specificity observed in the family has been reported for the *H. vulgare* Exo1 (7,19-21), or the β-glucosidases from *Thermotoga neapolitana* (8) and *Kluyveromyces marxianus* (9). However, the high variety in structure and composition found among the different enzymes make it difficult to extrapolate general rules explaining function, and a clear classification of different subfamilies is still needed.

A proper classification of GH3 glycosidases may require extensive biochemical and structural characterization of new enzymes. In this context, nature provides an inexhaustible reservoir from which enzymes can be isolated (22), because they are continuously changing and evolving as a consequence of natural processes of selection. Genomics and metagenomics have made accessible such enormous reserve of uncharacterized enzymes. Thus, we and others have recently taken advantage of sequencing and extensive screening technologies to develop enzyme discovery strategies and to identify microbial enzymes with improved and unusual activities and specificities (23-25), as well as distinct active site architectures and substrate preferences relative to other structurally characterized enzymes (26). These elegant studies demonstrated that nature contains proteins with novel and/or altered sequences and protein structures, the analysis of which represents one of the major challenges in postgenomic biology (27).

Here, activity screening of a metagenomic library created from rumen fluid led us to the isolation of a novel β-glycosidase, $GlyA_1$, which was assigned to the GH3 family. Detailed biochemical characterization of the new enzyme revealed its substrate specificity, whereas its sequence and crystal structure analysis revealed a novel permuted domain topology, defining a new subgroup within GH3 family. The enzyme contains an additional C-terminal domain, previously unidentified, its molecular flexibility being explored by small-angle X-ray scattering (SAXS) analysis. The structural and biochemical analysis of the $GlyA_1$ hydrolase presented in this study sheds new light on comparative catalysis and evolutionary model studies as well as phylogenetic relationships.

# RESULTS

*Library screening*- A subset of 14,000 clones from resident microbial communities of strained ruminal fluid (SRF) collected from rumen-fistulated, non-lactating Holstein cows (28) was screened for its ability to hydrolyse *p*-nitrophenyl-β-D-glucoside (*p*NPβGlc) and *p*-nitrophenyl-β-D-cellobioside (*p*NPβCel). We identified a positive clone (designated SRF4) being highly active against both substrates. The fosmid with insert SRF4 (38,710 bp; G+C 41.89%) was fully sequenced. A gene herein designated as *gly*$A_1$ encoding a potential GH3 β-glycosidase ($GlyA_1$) was identified out of the 38 distinct genes on the hit fosmid. The deduced molecular mass and estimated *p*I value were 101,849 Da and 4.86, respectively. This 921-amino acid-long putative protein exhibited a maximum amino acid sequence identity of 59% to a similar protein in public databases (with a top hit EDO57841.1 from *Clostridium* sp.). A search of oligonucleotide patterns against the GOHTAM database (29) and TBLASTX analysis revealed compositional similarities between the DNA fragment (38,710 bp) containing the gene for $GlyA_1$ with genomic sequences of *Eubacterium,*

*Butyrivibrio* and *Coprococcus* spp. BLASTN revealed similarities of short DNA fragments to *Prevotella* and *Paenibacillus* spp. BLASTX (search by translated DNA sequences) showed similarity to glycosidases of unknown *Clostridia* (phylum *Firmicutes)*. BLASTP search with identified protein sequences showed good matches for many of them against corresponding proteins in *Eubacterium*, *Prevotella*, members of *Lachnospiraceae*, *Clostridium*, *Ruminococcus* and *Bacteroides*. Most likely, $GlyA_1$ has thus its origin in the phylum *Firmicutes*, the presence of a phage gene may however indicate a horizontal gene transfer of the carbohydrate metabolism genes from *Firmicutes* to *Bacteroidetes*. Those microbes are known to be abundant in the ruminal environment and are thought to play key roles in the breakdown of proteins and carbohydrate polymers (30,31)

*Biochemical characterization of $GlyA_1$-* The gene encoding putative GH3 β-glycosidase ($GlyA_1$) was cloned, expressed in *E. coli* BL21 (DE3) and purified. The hydrolytic activity was analysed using 18 synthetic model *p*-nitrophenyl (*p*NP) derivatives with different sugars as well as a series of 11 additional oligosaccharides. Their specific activities (units/g protein) (Table 1) and the half-saturation (Michaelis) coefficient ($K_m$), the catalytic rate constant ($k_{cat}$), and the catalytic efficiency ($k_{cat}/K_m$) values (Table 2), were determined. As shown in Table 1 activity was confirmed for 18 substrates that revealed that $GlyA_1$ is a GH3 member with clear β-glucosidase and β-xylosidase activities, but also possessing β-galactosidase, β-fucosidase, α-arabinofuranosidase and α-arabinopyranosidase activities at low level, in this order (Table 1). The activity towards *p*NP-N-acetyl-β-D-glucosaminide (*p*NPGlcNAc) and *p*NP-N-acetyl-β-D-galactosaminide (*p*NPGalNAc) was below detection limit and thus the enzyme does not have β-N-acetyl-glucosaminidase nor β-N-acetyl-galactosaminidase activity. As shown in Table 2, in terms of catalytic efficiencies, *p*NPβCel was the preferred substrate, mainly due to the higher affinity for this substrate as compared to other *p*NP sugars. The purified recombinant hydrolase were also assayed for their activities toward different polymeric substrates. By meaning of specific activity determination, $GlyA_1$ hydrolysed all short cello- and xylo-oligosaccharides tested (degree of polymerisation [DP] from 2 to 5), with longer substrates being slightly preferred (Table 1). The catalytic efficiencies ($k_{cat}/K_m$) while using the non-activated substrates cellobiose and xylobiose were lower than those found for *p*NPβCel and *p*NP-β-xylobiose (*p*NPβXylb), respectively, mainly due to a significant decrease of $k_{cat}$ value for the natural disaccharides (Table 2). A comparison of kinetic parameters using the natural substrates xylobiose and cellobiose and the synthetic *p*NPβXylb and *p*NPβCel substrates confirmed the about 2-fold higher affinity for oligosaccharides containing β-linked glucosyl *vs* xylosyl substrates. Opposite, the affinities for the monosaccharides *p*NPβGlc and *p*NPβXyl were essentially similar, suggesting that affinity constrains are higher as the size of the oligosaccharides increases. However, due to the differences in $k_{cat}$ values no major differences in catalytic performance were observed when comparing β-Xyl and β-Glc-containing sugars. The catalytic performance ($k_{cat}/K_M$) found for other substrates is from low to very low mainly due to lower catalytic rates. The enzyme exhibited also activity against lichenan, suggesting that is able to hydrolyze substrates with mixed β-1,3/1,4 linkages. No activity was detected using avicel or filter paper, as well as towards substrates without β-1,4 linkages such as β-1,3 glucan, or mixed β-1,3/1,6 linkages like laminarin. Accordingly, the enzyme showed a clear preference for short cello-oligosaccharide substrates, which may likely be produced in natural settings from the cellulose components of plant cell walls due to the action of glucanases in the ruminal fluid. Other substrates such as gentibiose (containing D-glucoses joined by a β-1,6 linkage) and sophorose (or 2-O-β-D-glucopyranosyl-α-D-glucose) were also hydrolysed to a similar extend as cellobiose and xylobiose. The optimum activity for $GlyA_1$ was observed within a mesophilic range (45-65 ºC), and within a neutral or slightly acid pH (6.0-7.0), being most active at 55ºC and pH close to 6.5 (Figure 1).

*Biochemical characterization of $GlyA_1$-ΔCt-* A mutant containing a missing C-terminal region, herein referred $GlyA_1$-ΔCt, was created in the vector pQE80L. After purification, activity was determined for the 18 sugars being hydrolysed by the wild type enzyme, so that effect of the C-terminal region was tested. As shown in Table 1, the specific activity of the mutant was from 2 to 18.4-fold lower than that of the wild type, suggesting the importance of this region in the overall activity of the enzyme. The negative effect of the elimination of the C-terminal domain (compared to the full-length protein) was most notable for the hydrolysis of sugars containing β-glucose (from 11.3 to 17.1-fold activity reduction) as compared to those containing β-xylose (from 4.6 to 5.5-fold lower activity).

*Crystal structure determination*- Preliminary crystals from the wild-type $GlyA_1$ were obtained after more than 3 months with PEG3350 as the precipitant, and were cryoprotected into 25% D-glucose to obtain the complex with this sugar. The structure was solved by molecular replacement using the domains from *T. neapolitana* β-glucosidase as independent search models. Refinement and analysis of electron density maps allowed modelling of the chain containing residues 3-798, but did not show any density to build the C-terminal segment 800-921, suggesting a putative cleavage of this region in the slow crystallization step. The low numbers of crystals impeded analysis of the intact protein by mass spectrometry, but SDS-PAGE analysis of protein solution samples revealed the presence of two bands after incubation at room temperature or treatment with proteases. Therefore, the sample was incubated with subtilisin previously to the crystallization step, which accelerated formation of many good quality crystals, under similar conditions and with the same space group. These crystals were cryoprotected into 20% glycerol and this molecule was found bound at the active site. Furthermore, crystals from a truncated construct containing residues 2-799 ($GlyA_1$-ΔCt) grew also in a week from ammonium sulphate as precipitant and, despite having different shape, yielded the same cell and space group, which is consistent with the hypothesis that the wild-type sample was cleaved. These crystals were used to obtain the complexes with D-xylose and D-galactose. Many attempts done to crystallize the complete enzyme were unsuccessful. Also, a construct with residues 800-921, containing the isolated C-terminal region ($GlyA_1$-Ct), failed to crystallize. Crystallographic data and refinement statistics for the four structures here presented are given in Table 3.

*The permuted domain topology of $GlyA_1$*- The first solved structure from barley β-D-glucan glucohydrolase (7) showed the core structure common to GH3 enzymes, composed of an N-terminal $(\alpha/\beta)_8$ barrel domain 1 linked to an $(\alpha/\beta)_6$-sandwich domain 2 (Figure 2a), both of them providing residues that make up the active site. The later reported structures from *T. neapolitana* (8), *Trichoderma reesei* (12), *Aspergillus* (13,14) and *Listeria innocua* (18) β-glucosidases, and a β-glucosidase isolated from soil compost (32), showed the presence of an additional fibronectin type III (FnIII) domain (also designated fibronectin like domain, or FLD), located at the C-terminus. This three-domain arrangement is shared by other reported β−glucosidases from *K. marxianus* (9) and *Streptomyces venezuelae* (11) that also contain an additional PA14 domain inserted within the same loop of their $(\alpha/\beta)_6$-sandwich, although both are arranged in a different orientation. Moreover, the structure of the *Pseudoalteromonas* sp. exo-1,3/1,4-β-glucanase has been reported to have a C-terminal domain attached to the core structure, structurally related to family 30 carbohydrate-binding modules (CBM30) although its function is unknown (10). To expand even more this diverse landscape, $GlyA_1$ presents a novel structural arrangement showing permuted sequence and topology, in which the $(\alpha/\beta)_6$ sandwich (previous domain 2) is located at the N-terminus and the FnIII domain is sequentially inserted between this and the $(\alpha/\beta)_8$ barrel (Figure 2a). Additionally, a 120-residues segment is attached to the C-terminus most surely folded into an additional domain.

Figure 2b displays the 3D structure of the solved 3-798 region of $GlyA_1$, which present overall dimensions of 85 x 65 x 45 Å. The N-terminal $(\alpha/\beta)_6$-sandwich domain (red, residues 10-219) is followed by the FnIII domain (beige residues 278-419) and the $(\alpha/\beta)_8$ barrel domain (green, residues 468-780). Two long segments connect the three domains (grey). Linker 1 (residues 220-277) and half of linker 2 (residues 411-443) are tightly wrapped over the core structure, while the rest of linker 2 (444-467) forms an extended arm that clasps the $(\alpha/\beta)_8$ barrel. Finally, the regions at the beginning and the end of the chain are making a two-stranded β-sheet that laces the core structure at the top.

Comparative analysis using the Dali (33) server revealed that the $GlyA_1$ $(\alpha/\beta)_6$-sandwich domain, containing the catalytic acid/base residue Glu143, superimposes onto the corresponding domain from the *T. neapolitana* β-glucosidase, with a root-mean-square deviation (rmsd) of 1.6 Å for 202 equivalent Cα positions (39% sequence identity). The same comparison with the other structurally known GH3 gives deviations in the range 2-2.5 Å (20-25% sequence identity). The FnIII domain seems more structurally conserved along the GH3 family, the $GlyA_1$ being most similar to those in the β-glucosidases from *T. neapolitana*, with rmsd=1.5 Å (122 residues, 39% identity), and *K. marxianus*, with rmsd=1.6 Å (123 residues, 32% identity), but the same analysis gives values in the range 1.8-1.9 Å (21-28% sequence identity) against the other GH3 enzymes containing this domain. Finally, the $(\alpha/\beta)_8$ barrel, which contains the nucleophile Asp709, is most similar to the corresponding domain in the β-glucosidases from *T. neapolitana* (rmsd=1.6 Å, 285 residues, 39% identity), *K. marxianus*, (rmsd=1.5 Å , 276 residues, 34% identity), *Str. venezuelae* (rmsd=1.7

Å, 278 residues, 32% identity), and *Trichoderma reesei* (rmsd=2.0 Å, 278 residues, 27% identity). Equally to $GlyA_1$, all these domains present a deviation from the canonical $(\alpha/\beta)_8$ barrel topology, which was first observed in the *T. neapolitana* β-glucosidase. Thus, their first α–helix of the eight β–α motifs is missing, which has the consequence of making strand β2 reversed and antiparallel with the other seven strands. The different deviation from the canonical topology found at this domain is consistent with the higher deviations found in the structural comparison of $GlyA_1$ to other GH3 enzymes, in the range 2.5-3 A (16-20% identity).

Interestingly, the $GlyA_1$ core is structurally rather conserved with known β-glucosidases with equivalent domains architecture (Figure 2c). The superposition of the *T. neapolitana* β-glucosidase onto the structure of $GlyA_1$ here reported shows small differences in the orientation of some of the helices (Figure 2d). The main difference is the long arm that links the FnIII to the $(\alpha/\beta)_8$ domain in $GlyA_1$, which is missing in *T. neapolitana* β-glucosidase. There are also significant differences in the loops surrounding the active site both, in length and orientation, which must be related to the different substrate specificity, as commented below.

*The architecture of the active site*- The active site of $GlyA_1$ is located at the molecular surface, at the interface between the $(\alpha/\beta)_8$ barrel domain, which provides the nucleophile Asp709, and the $(\alpha/\beta)_6$-sandwich domain, contributing with the Glu143 acid/base catalyst (Figure 3a). The participation of Asp709 in substrate hydrolysis was confirmed by site-directed mutagenesis (D709A) in $GlyA_1$ and $GlyA_1$-ΔCt, as $K_m$ and $k_{cat}$ values could not be determined from the data obtained due to the activity value being below detection limit. It is a pocket of 12 Å deep with a narrow entrance 4-6 Å wide. A detailed structural comparison with the *T. neapolitana* β-glucosidase (Figure 3a) reveals the main differences in loop conformation observed around the active site that are responsible of making a deeper catalytic pocket in $GlyA_1$. First, loop β7–α7 of the $(\alpha/\beta)_8$ barrel, following the nucleophile Asp709 (residues 711-726), has an 11-residues insertion that extends away from the pocket and interacts with the long segment linking the FnIII domain to the barrel, which is missing in the *T. neapolitana* β-glucosidase. Here, Arg717 makes an ion-pair with Glu447 at the small helix located in the middle of the extended linker, which helps in stabilizing this region. An important feature of this β7–α7 loop is the presence of Trp711, close to the nucleophile Asp709, which protrudes from the surface and delineates a narrow catalytic pocket. Moreover, and despite loop β3–α3 (residues 536-550) being shorter in $GlyA_1$, Arg538 clearly bulges into the pocket contributing to constrict it even more.

With respect to the $(\alpha/\beta)_6$-sandwich, and similarly to that observed in *T. neapolitana* β-glucosidase, this domain is shaping the active site by means of two loops, residues 139-152 containing the acid/base catalyst Glu143, and residues 100-113 enclosing Trp106 that clearly projects into the catalytic pocket. Interestingly, the last loop is markedly flexible as it is deduced from the fact that it could only be fully traced in the ligand-free crystal, containing only glycerol in the active site, and in the galactose-soaked crystal of the truncated form. In contrast, the crystals of the full-length and truncated forms, respectively soaked into glucose and xylose, showed poor density that precluded tracing residues 104-107. Furthermore, the traced loops showed significant conformational changes in the different crystals at Trp111, coupled to a change in Phe147 from the adjacent 139-152 loop (Figure 3a), reinforcing its intrinsic mobility. The loop equivalent to 100-113, which is highly variable within GH3 enzymes, was proposed to be involved in recognition of large substrates from the crystal structure of *T. neapolitana* β-glucosidase, which showed some disorder that precluded tracing of a segment equivalent to that non-observed in some $GlyA_1$ crystals. Noteworthy, the non-visible region of *T. neapolitana* β-glucosidase includes Trp420 that, consequently, may be defining additional binding subsites, similarly to Trp106. However, the remaining sequence is not conserved, both Phe147 and Trp111 being unique to $GlyA_1$ and, therefore, the substrate recognition mode presented by the two enzymes to accommodate the substrate may be different.

Soaking with xylose and glucose showed a clear density indicating that both sugars occupy the catalytic pocket subsite -1 in a relaxed chair conformation (Figure 3b). This subsite is well conserved among known GH3 β-glucosidases and, with the exception of the acid base catalyst, is made up entirely by residues from the $(\alpha/\beta)_8$ barrel domain. Thus, residues from the loops emerging from the central β-strands are making a tight net of hydrogen bonds that accommodate the glycon with all its OH groups making at least two polar interactions. The glycon moiety is located by stacking to Trp710, and the acid base catalyst Glu143 and the nucleophile Asp709 interact with the O1 and O2 hydroxyls, as it is expected in GH enzymes. The other residues making subsite -1 are Asp532, Arg597, Lys630, His631, Arg641 and

Tyr677. Xylose and glucose are bound in an identical position and the glycerol molecules observed in the ligand free crystals are mimicking the positions occupied by C2, C3, C4 and C5 from both sugars. The additional polar interaction made by the glucose O6 hydroxyl appears consistent with the higher affinity observed in $GlyA_1$ towards glucosides as compared to xylosides. Thus, as shown in Table 2, the affinity for cellobiose ($K_m$=2.4±0.3 mM) was approximately 2-fold higher than that for xylobiose ($K_m$=4.7±0.2 mM). Interestingly, soaking of crystals with galactose showed that this sugar displays a semi-chair conformation at subsite -1, by flattening of the C4 atom that has the axial hydroxyl substituent (Figure 3b, inset). In this way, galactose is accommodated by essentially the same polar interactions observed in the glucose complex, therefore explaining the activity of the enzyme on β-galactosides. However, the energy cost of getting the substrate ring distortion is reflected by the lower β-galactosidase activity, as given in Tables 1 and 2. Accordingly, the low β-fucosidase and α-arabinosidase activities must reflect some degree of deviation from the glucose-binding pattern, through ring distortion and/or loss of polar interactions, but in any case, the plasticity of the catalytic site provides a notable capacity of $GlyA_1$ to accept different sugars (from high, to low and very low specificity).

As said before, and in contrast to that observed in *T. neapolitana* β-glucosidase that presents an active site opened to the solvent with only subsite -1 being defined, more subsites are apparent in $GlyA_1$. To delineate a putative +1 subsite, we modelled the position of the non-hydrolysable substrate analogs thiocellobiose and thiogentibiose, by structural superimposition on the previously reported experimental barley complexes (34). As it is shown in Figure 3c, Trp106 and Trp711 define a hydrophobic patch that may allocate the oligosaccharides at a putative subsite +1, leaving a range of possible ring orientations compatible with the observed activity of $GlyA_1$ against differently β-linked bioses, as given in Table 1. Also, the long chain of Arg538, protruding at the catalytic pocket as said above, is in good position to stabilize the sugar unit by making hydrogen bonds to one or possibly two of its hydroxyl groups. The important contribution of subsite +1 to GlyA1 substrate binding efficiency (both glucosides and xylosides) is manifested by the lower $K_m$ with *p*NPβCel compared to *p*NPβGlc and by the lower $K_m$ with *p*NPβXylb compared to *p*NP $\beta$ Xyl (Table 2).

Furthermore, inspection of the molecular surface of the active site cavity shown in Figure 3d suggests the possible existence of additional subsites, which is illustrated by several β−1,4/1,3−linked oligosaccharides that have been modelled at the active site: a glucotetraose (green), a Glc-4Glc-3Glc-4Glc chain (purple) and a Glc-4Glc-4Glc-3Glc (yellow). These sugars have been docked by superimposition of their non-reducing units onto the observed glucose at the $GlyA_1$ complex. The hydrophobic patch defined by Trp106 and Trp711 may fit the oligosaccharides at subsites +1 and +2, and the long side-chain of Lys723 seems available to make polar interactions with the hydroxyl groups defining a possible subsite +3. The putative existence of at least three subsites in $GlyA_1$ active site would be in agreement with the tendency of an increased activity against longer cello- and xylo oligosaccharides, given in Table 1. Also, the tendency of an increased activity against longer cello- and xylo oligosaccharides given in Table 1 suggests interactions at more distal positions and, therefore, the possibility of additional subsites. On the other hand, the shape of the active site seems compatible with the mixed β−1,4/1,3−links of the modelled tetrasaccharides, explaining therefore the observed activity on the medium-size polymer lichenan.

The comparison of $GlyA_1$-Glc structure with those reported for *T. reesei* β-glucosidase (12) and barley β-D-glucan glucohydrolase complexed with thiocellobiose (7), shown in Figure 3e, displays the different hydrophobic platforms found at each active site. The barley β-D-glucan glucohydrolase structure showed a narrow channel with the glucose tightly arranged at subsite +1, being sandwiched between Trp286 and Trp434 side chains. In contrast, the $GlyA_1$ Trp711 is perpendicular and oriented similarly to Trp37 found in *T. reesei* β-glucosidase, although both residues are provided by different loops from the $(\alpha/\beta)_8$ barrel domain. At the opposite face, GlyA1 Trp106 is structurally equivalent to Tyr443 and Trp434 from the barley and *T. reesei* enzymes, although all of them come from different loops within the $(\alpha/\beta)_6$-sandwich domain. Interestingly, other enzymes present an aromatic residue in a position identical to Trp106, but they are provided by the PA14 domain, Phe508 in the case of the *K. marxianus* β−glucosidase, or by a long loop coming from the other subunit, Tyr583 in the case of the *Listeria innocua* β-glucosidase dimer (18) (not shown). This feature illustrates that these highly diverse enzymes have evolved common topology and molecular mechanisms and, yet, the precise structural differences behind that regulate specificity.

*SAXS analysis of GlyA$_1$*- Due to the unfeasibility in crystallizing the full-length GlyA$_1$, we explored its overall flexibility and putative shape in solution by SAXS experiments. Thus, we compared the molecular descriptors of the complete construct with respect to the truncated construct GlyA$_1$-ΔCt, lacking the C-terminal domain. For this purpose, several solutions with varying concentrations were measured for each sample, and their scattering curves were merged to extrapolate idealized data. Analysis of the scattering curves show a good fit to the Guinier approximation, which indicates that the samples are not aggregated. Also, the calculated radii of gyration (Rg) are consistent across the range of measured concentrations. Then, the overall size descriptors can be properly determined for each construct.

First of all, the calculated molecular masses from both samples are close to the expected values (Table 4), indicating the presence of monomers, and also a 15 kD higher mass in the complete protein, which excludes proteolysis of the analysed sample in the short time of the experiment. On the other hand, the Rg and the maximum distance (Dmax) for the complete protein are only slightly higher than the truncated protein, which may be indicating that the extra C-terminal domain is not too extended from the core structure. In support of this hypothesis, the pair-wise distance distribution function P(r) calculated for both constructs shows a similar unimodal pattern consistent with a single domain protein in both cases. Furthermore, the analysis of the scattering function by the Kratki plots is consistent with the expected profile for a folded protein with a clear peak, in contrast what is observed in multidomain proteins with flexible linkers that present several peaks or smoother profiles. Consequently, we do not observe in the data calculated from the complete protein any of the signs that may be indicative of molecular flexibility, i.e, large Rg and Dmax, absence of correlation in the P(r) function or smooth Kratky plots. Therefore, SAXS analysis appears consistent with a compact overall shape of the complete GlyA$_1$, in which the extra C-terminal region would not define a marked separate or flexible domain but, rather, it could be folded over the core three-domain structure.

To test the feasibility of this hypothesis, *ab initio* models were generated for complete GlyA$_1$ from SAXS data. First, two models of the last 120 residues (GlyA$_1$-Ct) were obtained, as explained in the experimental section, both showing an overall β-sandwich topology. This topology is related to carbohydrate-binding domains within families CBM6 and CBM35, to which GlyA$_1$-Ct presents 15-20% sequence identity, although the equivalent carbohydrate binding motifs, typically clusters of conserved aromatic residues, are not evident in its surface. Then, three runs of CORAL were computed by considering the experimental structure of the truncated protein and each of the two models. The six models obtained are shown in Figure 4. Analysis of these models reveals that all of them cluster around a reduced area that would locate the C-terminal region relatively distant from the catalytic pocket but quite near the mobile loop (residues 100-113). Overall, these models are consistent with the hypothesis proposed above, suggesting that GlyA$_1$-Ct may be somewhat packed between the two domains making the core structure and, interestingly, with a putative linker somehow exposed to solvent. This feature might possibly explain the proteolysis observed in the complete protein.

*GlyA$_1$ phylogenetic analysis*- Our structural analysis illustrated that the permuted domain architecture of GlyA$_1$ keeps the location of the active site at the interface between the $(\alpha/\beta)_8$ barrel and the $(\alpha/\beta)_6$-sandwich domains. As mentioned above, N-acetyl-glucosaminidases are built by a single domain, with its $(\alpha/\beta)_8$ barrel holding both, the nucleophile and acid/base catalyst. Interestingly, the *Bacillus subtilis* NagZ shows the two-domain composition but still keeps the catalytic residues at the $(\alpha/\beta)_8$ barrel (16). Therefore, this domain may be considered as the characteristic signature of GH3 enzymes. In order to examine the phylogenetic positioning of β-glucosidases with inverted topology (represented by GlyA$_1$) within the GH3 family, we have carried out a phylogenetic analysis based on the sequence of its $(\alpha/\beta)_8$ barrel domain (ABB in this analysis). Sequences representative for each of the domain architectures found in the GH3 domain were selected (details in Experimental procedures). The five topologies selected for this study are ABB, ABB-ABS, ABB-ABS-FLD, ABS-FLD-ABB and ABB-ABS(PA14)-FLD (ABS $(\alpha/\beta)_6$-sandwich; FLD fibronectin-like type III domain). The resulting phylogenetic tree given in Figure 5 shows apparent correlation between ABB sequence divergence and domain architecture. Most single domain sequences (ABB) cluster together and correspond to N-acetyl-glucosaminidases (salmon area of the tree). Insertion of the ABS module is associated to three different nodes (a, b and c in Figure 5). Insertion at node (a) was not accompanied by a significant divergence in the ABB sequence since both ABB and ABB-ABS architectures appear mixed at this node. In fact, these ABB-ABS sequences also

correspond to N-acetyl-glucosaminidases, and crystallographic data of *B. subitilis* NagZ show that the two modules are quite independent from a structural point of view. ABS insertion at nodes (b) and (c) would correspond to the divergence of GH3 enzymes giving rise to other activities, mainly β-glucosidase. Within node (c), other modules (FLD and PA14) were appended after ABS. At node (b), fusion of C-terminal FLD seems to occur close to ABS addition since most sequences contain both modules. $GlyA_1$ and the other GH3 enzymes with inverted topology arose within this cluster. The phylogenetic analysis shows that the inverted topology is predominantly found in Firmicutes, although it is also present in at least another phylum (Actinobacteria) and even Archaea. Furthemore, it appears clearly associated to enzymes belonging to bacteria dwelling in the digestive tract of animals.

## DISCUSSION

In the present work, a functional metagenome library analysis was used to identify a β-glycosidase from a plant polymer-degrading microorganism populating the rumen of a dairy cow. The enzyme most likely originated from the genome of a representative of Firmicutes phylum known to be abundant in the ruminal environment (30,31).

The structural and biochemical analysis of the $GlyA_1$ hydrolase presented in this study sheds new light on the mechanisms of the catalysis and evolutionary patterns of the GH3 family. Our data demonstrated that $GlyA_1$ has a permuted domain topology. It is well documented that the formation of new domain combinations is an important mechanism in protein evolution. The major molecular mechanism that leads to multi-domain proteins and novel combinations is non-homologous recombination, sometimes referred to as 'domain shuffling'. This may cause recombination of domains in order to form different domain architectures. Proteins with the same series of domains or domain architecture are related by descent (i.e. evolved from one common ancestor), and tend to have the same function (35), which is rarely the case if domain order is switched. Indeed, a detailed analysis of the structures of proteins containing Rossmann fold domains demonstrated that the N- to C terminal order of the domains is conserved because the proteins have descended from a common ancestor. For pairs of proteins in the PDB in which the order is reversed, the interface and functional relationships of the domains are altered (36). This was also proved in this study that revealed that altered domain architecture in GH3 mostly evolved from a distinct ecological niche, most likely from digestive tracts, including that of the ruminants. Also, the substrate specificity of the $GlyA_1$ protein is markedly different to that of reported GH3 members. Indeed, $GlyA_1$ is a uncommon multi-functional GH3 with β-glucosidase, β-xylosidase, β-galactosidase, β-fucosidase, α-arabinofuranosidase, α-arabinopyranosidase and lichenase co-activities, with the ability to degrade β-1,2-, β-1,3-, β-1,4-, and β-1,6-glucobioses.

From an ecological point of view, the rumen compartment provides stable and favorable conditions for microbial growth that is also permanently exposed to plant biomass; for this reason it contains specialized microorganisms that are permanently competing or collaborating for the degradation of the plant fibers. The data herein presented suggests that this factor, namely the high exposure to plant biomass, which is less common in other habitats, may be a strong force driving the establishment of gut microbiota with GH3 protein with permuted structures that may provide ecological advantages. Indeed, the permuted domain topology may confer the protein different functionalities such as the ability to expand the pool of biomass-like substrates being hydrolyzed. Overall, our results (analysis of oligonucleotide pattern and phylogenetic tree) strongly suggest that $GlyA_1$ and related GH3 enzymes with inverted topology emerged in Firmicutes, where their presence is rather frequent, being transferred by horizontal gene transfer to bacteria from other phyla and even to other kingdom (archaea). It is well documented that these wide range gene transfer events take place at high frequency in the rumen (37,38). Probably, $GlyA_1$ topology arose from a sequence encoding a GH3 enzyme with ABB-ABS-FLD domain architecture, by gene inversion. Although the inversion surely rendered a nonfunctional gene, further mutations that would restore some sort of glycolytic activity would be strongly favored by selective pressure.

Structural analysis illustrates the permuted domain composition of $GlyA_1$ which is composed by an N-terminal $(\alpha/\beta)_6$-sandwich domain, followed by the FnIII domain and the $(\alpha/\beta)_8$ barrel domain. Based on sequence data a C-terminal domain was expected after the $(\alpha/\beta)_8$ barrel domain. However, attempts to crystallize the C-terminal region of the protein were unsuccessful, and its functional role was unclear. Biochemical characterization of the $GlyA_1$ and $GlyA_1$-ΔCt proteins revealed that the C-terminal domain do not affect the overall substrate profile of the protein, but rather it affects the catalytic performance,

which is significantly lower in the truncated $GlyA_1$-ΔCt protein. This suggests that most likely the C-terminal domain may not have a direct role in substrate binding but, still, it might disturb the dynamics of the proximate mobile loop (residues 100-113), which seems directly involved in catalysis.

According to available structure-prediction tools, this C-terminal region is expected to adopt a lectin-like topology, related to the CBM6/CBM35 domains. However, it does not seem an obvious carbohydrate-binding domain and, in fact, binding to xylan, cellulose and barley glucan was not observed by affinity gel electrophoresis assays (not shown). Nevertheless, although its involvement in binding small substrates does not seem apparent, this domain might be playing a role in positioning or locating the enzyme to distal positions of a yet unknown polymeric substrate, by recognizing specific but still unidentified substitutions. Alternatively, it could play a role in keeping the enzyme attached to the cell surface, facilitating the intake of its products and conferring the bacteria an advantage over competing organisms. Interestingly, the analysis of the $GlyA_1$-Ct homologous sequences shows that these domains are attached to GH3 β-glucosidases from ruminal environment, this feature pointing at a possible function related to this ecosystem. However, its presence is not related to the permuted domain topology, as only half of the sequences included in the $GlyA_1$ cluster (Figure 5) contain segments equivalent to $GlyA_1$-Ct.

In conclusion, the analysis of $GlyA_1$ here presented uncovers new features of GH3 enzymes and provides a template for a novel subfamily including members with permuted domain topology. It also allows picturing the $GlyA_1$ active site architecture and the molecular basis of its substrate specificity. More work is needed to have a complete picture of the intricate molecular mechanisms that these highly diverse enzymes have evolved to tailor specificity. It will contribute to improve our knowledge about enzymatic carbohydrate degradation and open up new avenues for biocatalysis.

## EXPERIMENTAL PROCEDURES

*Reagents and strains*- Chemicals and biochemicals were purchased from Fluka-Aldrich-Sigma Chemical Co. (St Louis, MO, USA) and Megazyme (Bray, Ireland) and were of p.a. (pro-analysis) quality. The oligonucleotides used for DNA amplification were synthesised by Sigma Genosys Ltd. (Pampisford, Cambs, UK). The *Escherichia coli* Rosetta2 (Novagen, Darmstadt, Germany) for cloning and expression of wild type protein and genetic constructs in pQE80L vector were cultured and maintained according to the recommendations of the suppliers.

*Metagenomic library screening and positive-insert sequencing*- A pCC1FOS fosmid metagenomic library created from microbial communities from SRF of rumen-fistulated, non-lactating Holstein cows, was used. The construction and characteristics of the library was described previously (28). A subset of 14,000 clones were plated onto large (22.5 x 22.5 cm) Petri plates with Luria Bertani (LB) agar containing chloramphenicol (12.5 μg/ml) and an arabinose-containing induction solution (Epicentre Biotechnologies; WI, USA) at a concentration (0.01% *w*/*v*) recommended by the supplier to induce a high fosmid copy number. After overnight incubation at 37 ºC, the clones were screened for the ability to hydrolyse *p*NPβGlc and *p*NPβCel. For screens, the plates (22 x 22 cm; each containing 2,304 clones) were covered with an agar buffered substrate solution (40 ml of 50 mM sodium acetate buffer, pH 5.6, 0.4% w/v agar and 5 mg/ml of *p*NPβGlc and *p*NPβCel as substrates). Positive clones were detected by the formation of a yellow colour. One positive clone, herein designated as SRF4 was selected, and its DNA insert fully sequenced with a Roche 454 GS FLX Ti sequencer (454 Life Sciences, Branford, CT, USA) at Life Sequencing S.L (Valencia, Spain), and predicted genes were identified as described previously (28).

*Cloning of $glyA_1$ and genetic constructs in pQE80L plasmid*- The full coding sequence of $GlyA_1$ (residues 2-921) and a deleted version (residues 2-799) lacking the C-terminal domain ($GlyA_1$-ΔCt) were amplified by PCR with 4GF (CACGAGCTCAATATTGAAAAAGTGATACT TGATTGG) as forward oligonucleotide and 4GR1 (AGCCGTCGACTTACTGCTGCTTTTTAAACT CTATTCG) or 4GR2 (AGCCGTCGACTTACACTCTTCCTGCTATCT CAACC) as reverse oligonucleotides, respectively. The SRF4 fosmid was used as the template. The PCR conditions were as follows: 95 °C for 120 s, followed by 30 cycles of 95 °C for 30s, 55 °C for 45 s and 72 °C for 120 s, with a final annealing at 72 °C for 500 s. The PCR products were analysed, and agarose gel-purified using the Mini Elute Gel Purification Kit (Qiagen, Hilden, Germany). The PCR products were digested with SacI/SalI and cloned in vector pQE80L to generate plasmids $GlyA_1$-pQE and $GlyA_1$ΔCt-pQE, respectively. The coding sequence of the C-terminal domain ($GlyA_1$-Ct, residues 800-921) was amplified with oligonucleotides CT1F

(CACGAGCTCATAGAAGAGGATGCATTCGA TATAG) and 4GR1, and cloned in the SacI/SalI sites of pQE80L (plasmids Ct-pQE). $GlyA_1$-pQE was used as a template to introduce the mutation D709A by PCR with primers M1 (TGGTGGGCTCAGGTTAATGACC) and M2 (G***G***CAGTCATCACAATACCCTTAAAGCC), as previously described (39). The coding region of the resulting plasmids was fully sequenced to check for the absence of undesired mutation. The *E. coli* strain Rosetta2 (Novogen, Darmstadt, Germany) was transformed with the selected plasmids; the clones were selected on LB agar supplemented with ampicillin (100 μg/ml) and chloramphenicol (68 μg/ml) and stored with 20% (*v/v*) glycerol at -80 °C.

*Site directed mutagenesis*- Mutation D709A was introduced into the corresponding pQE80L plasmids containing genes encoding $GlyA_1$ and $GlyA_1$-ΔCt, using the QuikChange II XL Mutagensis Kit from Agilent Technologies Inc (Santa Clara, CA, USA), with TGGTGGGCTCAGGTTAATGACC and GGCAGTCATCACAATACCCTTAAAGCC as forward and reverse oligonucleotides, respectively. The resulting variant plasmids were transferred into *E. coli* strain Rosetta2 (Novagen, Darmstadt, Germany) and selected on the LB agar supplemented with same antibiotics as parental plasmids.

*Gene expression and protein purification*- For the enzyme expression and purification of wild-type and mutant $GlyA_1$ and $GlyA_1$-ΔCt variants, as well as $GlyA_1$-Ct in pQE80L vector, a single colony (*E. coli* Rosetta2) was grown overnight at 37 ºC with shaking at 200 rpm in 100 ml of 2x-TY medium (1% yeast extract, 1.5% triptone, 0.5% NaCl) containing ampicillin (100 μg/ml) and chloramphenicol (68 μg/ml), in 1 L flask. Afterwards, 25 ml of this culture was used to inoculate 1 L of 2x-TY medium, which was then incubated to an $OD_{600nm}$ of ~0.6 (range from 0.55 to 0.75) at 37 ºC. Protein expression was induced by 0.9 mM isopropyl-β-D-galactopyranoside (IPTG) followed by incubation for 16 h at 16 ºC. The cells were harvested by centrifugation at 5000 × *g* for 15 min to yield 2-3 g/l of pellet (wet weight). The cell pellet was frozen at -80 ºC overnight, thawed and resuspended in 3 ml 20 mM phosphate buffer pH 7.4, 500 mM NaCl per gram of wet cells. Lysonase Bioprocessing Reagent (Novagen, Darmstadt, Germany) was then added (4 μl/g wet cells) and incubated for 30 min on ice with rotating mixing. The cell suspension was then sonicated for a total of 1.2 min and centrifuged at 15000 × *g* for 15 min at 4 ºC; the supernatant was retained. The $His_6$-tagged enzyme was purified at 4 °C after binding to a Ni-NTA His·Bind resin (Novagen, Darmstadt, Germany). The columns were pre-washed with 20 mM phosphate buffer pH 7.4, 500 mM NaCl and 50 mM imidazole, and the enzyme was eluted with the same buffer but containing 500 mM imidazole. The monitoring of the enzyme elution was performed by SDS-PAGE and/or activity measurements, using standard assays (see above). After elution, protein solution was extensively dialyzed with Tris 20 mM pH 7.5, 50 mM NaCl by ultra-filtration through low-adsorption hydrophilic 10,000 nominal molecular weight limit cutoff membranes (regenerated cellulose, Amicon), after which the protein was maintained at a concentration of 10 mg/ml; the protein stock solution was stored at -20 ºC until used in assays. The purity was assessed as >95% using SDS-PAGE, which was performed with 12% (v/v) polyacrylamide gels, using a Bio-Rad Mini Protean system. Prior to crystallization assays 2 mM dithiothreitol (DTT) was added.

*Biochemical assays*- Specific activity (units/g) and kinetic parameters ($K_m$ and $k_{cat}$) were firstly determined using *p*NP sugars (read at 405 nm) in 96-well plates, as previously described (28). *p*NP substrates tested included those containing: α-glucose (*p*NPαGlc), α-maltose (*p*NPαMal), β-glucose (*p*NPβGlc), β-cellobiose (*p*NPβCel), α-arabinofuranose (*p*NPαAraf), β-arabinopyranose (*p*NPβArap), α-xylose (*p*NPαXyl), β-xylose (*p*NPβXyl), β-xylobiose (*p*NPβXylb), α-fucose (*p*NPαFuc), α-rhamnose (*p*NPαRha), α-mannose (*p*NPαMan), β-mannose (*p*NPβMan), α-galactose (*p*NPαGal), β-galactose (*p*NPβGal), β-lactose (*p*NPβLac), N-acetyl-β-D-glucosaminide (*p*NPGlcNAc), and N-acetyl-β-D-galactosaminide (*p*NPGalNAc). For cello-oligosaccharides (DP from 2 to 5), gentibiose and sophorose, the level of released glucose was determined using a glucose oxidase kit (Sigma Chemical Co., St. Louis, MO, USA). The level of released xylose from xylo-oligosaccharides (DP from 2 to 5), was determined using the D-xylose assay kit from Megazyme (Bray, Ireland). Substrate specificity was investigated also using carboxymethylcellulose (CMC), lichenan, barley glucan, laminarin and avicel (all from Sigma Chemical Co., [St. Louis, MO, USA]) and filter paper (Whatman, England). Specific activity for all these sugars was quantified by measuring release of reducing sugars according to Miller (1959). For $K_m$ determinations, assay reactions were conducted by adding a protein concentration of 0.23 μM to an assay mixture containing from 0 to 30 mM sugar in 50 mM sodium acetate buffer, pH 5.6, $T$= 40 ºC. Total reaction volume was to 200 μl. For $k_{cat}$

determinations, under the same conditions, sugar concentration was set up to 2 times the $K_m$ value and the protein concentration from 0 to 0.23 μM. For specific activity determinations (units/g) a protein concentration of 0.23 μM and 10 mg/ml of the sugar or polysaccharide were used in 50 mM sodium acetate buffer pH 5.6, *T*= 40 ºC. The pH and temperature optima were determined in the range of pH 4.0–8.5 (50 mM Britton–Robinson buffer, BR) and 20–65 °C in assays containing a protein concentration of 0.23 μM and 10 mg/ml of *p*NPβGlc, which was used as standard substrate. BR buffer is a "universal" pH buffer used for the range pH 2 to pH 12. It consists of a mixture of 0.04 M $H_3BO_3$, 0.04 M $H_3PO_4$ and 0.04 M $CH_3COOH$ that has been titrated to the desired pH with 0.2 M NaOH. Optimal pH was measured at 40 ºC, and the optimal temperature was determined in the same buffer used in the kinetic assays. In all cases, absorbance was determined immediately after reagent and enzyme were mixed using a microplate reader every 1 min for a total time of 15 min (Synergy HT Multi-Mode Microplate Reader, BioTek). All reactions were performed in triplicate. One unit (U) of enzyme activity was defined as the amount of enzyme required to transform 1 μmol of substrate in 1 min under the assay conditions, with extinction coefficient as in (21). All values were corrected for non-enzymatic hydrolysis (background rate).

The protein concentration was determined spectrophotometrically (at 280 nm) using a BioTek EON microplate reader (Synergy HT Multi-Mode Microplate Reader - BioTek) according to extinction coefficient of the protein (108,485 $M^{-1}$ $cm^{-1}$) corresponding to its amino acid sequence (www.expasy.org/tools/protparam.html).

Note that the detection limit, using a microplate reader with a filter for 405 nm, for the yellow chromogen is about $1 \cdot 10^{-6}$ mol/L of *p*-nitrophenol. Since the concentration of substrate in the assay is ranging from 0 to 30 mM, it is expected that detection of the activity under our assay conditions is much above detection limit.

*Crystallization data collection and crystal structure determination*- Initial crystallization conditions for the complete $GlyA_1$ (10 mg/ml) were explored by high-throughput techniques with a NanoDrop robot (Innovadyne Technologies Inc.), using six different commercially screens: PACT and JCSG+ Suites from Qiagen; JBScreen Classic 1-4 from Jena Bioscience; and Index, Crystal Screen and SaltRx packages from Hampton Research. These assays were carried out using the sitting-drop vapour-diffusion method in MRC 96 well crystallization plates (Molecular Dimensions).

Elongated bars grew after three months in 20% polyethyleneglycol (PEG) 3350, 0.2 M ammonium sulphate, Bis-Tris pH 5.5. For data collection, crystals were cryoprotected in mother liquor supplemented with 25% (v/v) D-glucose before being cooled in liquid nitrogen. Diffraction data were collected at the German Electron Synchrotron (Hamburg, Germany). Diffraction images were processed with XDS (40) and scaled using Aimless from the CCP4 package (41) leading to space group $P2_12_12_1$. The structure was solved by molecular replacement using MOLREP (42) with reflections up to 2.5 Å resolution range and a Patterson radius of 54 Å. The template model was the β-glucosidase from *T. neapolitana* (PDB code 2X42), but the search was made in two steps. First, the region containing residues 2-315 was used for finding a partial solution. Then, another round of molecular replacement, with the region 321-721, was computed. Preliminary rigid-body refinement was carried out using REFMAC (43). Subsequently, several rounds of extensive model building with COOT (44) combined with automatic restrain refinement with flat bulk solvent correction and using maximum likelihood target features, led to a model covering residues 3 to 798. However, no density was found for the loop 103-108 or for the last 123 residues of the protein. At the latter stages, β-glucose, sulphate ions and water molecules were included in the model, which, combined with more rounds of restrained refinement, led to a final R-factor of 15.7 (Rfree 17.8). The free R-factor was calculated using a subset of 5% randomly selected structure-factor amplitudes that were excluded from automated refinement. Many attempts to reproduce and improve these crystals were unsuccessful, until *in situ* proteolysis of the sample with subtilisin was tried. Resulting crystals grew after 15 days in the same conditions but at pH 7.0, were cryoprotected in 20 % glycerol and showed the same space group and cell content. Then, the truncated $GlyA_1$-ΔCt construct (residues 1-798) was tested. Initial crystallization assays were accomplished as described above and several hits were obtained. Best crystals were grown in 2.0 M ammonium sulphate, 0.1 M Bis-Tris pH 5.5, and belonged to the same space group. The asymmetric unit contains a single molecule, with a Matthews's coefficient of 2.73 and a 54% solvent content within the cell.

Soaking experiments with D-xylose or D-galactose were performed with the truncated construct in mother liquor solution implemented with 5-50 mM ligand. Then, the crystals were flash frozen into liquid nitrogen using mother liquor plus 20% (v/v) glycerol or ethylene glycol as

cryoprotectants. The ligands were manually modelled into the electron density maps and were refined similarly to that described above. Although a mixture of α- and β- anomers may exist in solution, only the β-form of the monosaccharides was observed at the active site of the different complexes. For the docked glucotetraoses coordinates, not present in the Protein Data Bank, a model was built by the online carbohydrate-building program GLYCAM (45).

Many attempts to crystallize the C-terminal section of the protein using the available construct were unsuccessful and, thus, a model was built as explained in the next section. The figures were generated with PyMOL (46). The atomic coordinates have been deposited in the RCSB Protein Data Bank under the accession codes 5K6I, 5K6M, 5K6N, 5K6O.

*Small-angle X-ray scattering (SAXS) measurements*- $GlyA_1$ and $GlyA_1$-ΔCt stock solutions (10 mg/ml) were dialyzed against the same buffer (20 mM Tris-HCl pH 7.5, 50 mM NaCl, 2 mM DTT and 5% glycerol) for 18 h. SAXS measurements were performed at ESRF on beamline BM29, equipped with a Pilatus 1M detector. Each sample concentration, prepared by dilution of these stock solutions, was measured in 10 frames, 1 s exposure time per frame, at 4 °C, at a sample-to-detector distance of 2.867 m, using an X-ray wavelength of 0.991 Å. No radiation damage was observed during the measurements. The SAXS curves for buffer solutions were subtracted from the protein solution curves before analysis.

The scattering curves from six gradual concentrations, from 0.3 to 5 mg/ml, were scaled and averaged to obtain the I(q) function using ATSAS software package (47). The radius of gyration (Rg) for each protein was calculated by Guinier plot using the program PRIMUS, and the pair distribution function P(r) and the maximum particle size Dmax were obtained by the program GNOM. Then, POROD was used to calculate the excluded volume of the particle, as well as the molecular weight of each sample.

Several homology and threading modelling programs were tried to obtain a model of the last 123 residues of $GlyA_1$. All of them predicted a topology corresponding to carbohydrate-binding domains of families CBM6/CBM35, but differed in the length of the linker attaching this domain to the core protein. Finally, models obtained from Swiss-Model (48) and CPH-model (49) servers were used (templates from PDB entries 2W46 and 1UYX), each predicting a loop of 32 or 5 residues, respectively. Both entries share less than 20% identity with the C-terminal region of $GlyA_1$.

Subsequently, CORAL (47) was used for several rounds of two-domain rigid body fitting, using the $GlyA_1$-ΔCt coordinates and both templates, alternatively; linkers were built as dummy-atoms. The fit of the CORAL models to the SAXS experimental data was evaluated by the $\chi^2$ value calculated from the program CRYSOL (47).

*Sequence analysis and construction of a Neighbor-Joining tree*- The positioning of the sequence of the $GlyA_1$ $(\alpha/\beta)_8$ barrel domain was analyzed in a phylogenetic tree. The predicted protein sequences were aligned against the National Centre for Biotechnology Information non-redundant (NCBI nr) database using BLASTP algorithm. We downloaded all 27,499 GH3 sequences deposited in public databases. They were grouped within 5 different domain architectures: ABB (9,196), ABB_ABS (3,392), ABB_ABS_FLD (11,910), ABB_ABS_PA14_FLD (2,673) y ABS_FLD_ABB (328), where ABB, ABS, FLD and PA14 refer to $(\alpha/\beta)_8$ barrel domain, $(\alpha/\beta)_6$-sandwich, fibronectin-like type III domain and protective antigen PA14 domain, respectively. We discarded those sequences (848) from the ABB_ABS group lengthier than 700 amino acids, as they represent enzymes with unidentified domains downstream the ABS module. Subsequently, the sequence corresponding to the ABB domain was extracted from all of the 5 sub-groups. An additional filter was applied to remove ABB sequences with coverage lower than 60% of the consensus domain defined by Interpro or Pfam databases (i.e. with less than 200 amino acids). The final number of sequences was the following: ABB (8,109), ABB_ABS (2,312), ABB_ABS_FLD (7,335), ABB_ABS_PA14_FLD (1,664) y ABS_FLD_ABB (289). For each of the 5 sub-groups, redundant sequences (those sharing more than 50% identity) were eliminated to select sequences that belong to different taxonomic groups. Following this procedure the final selected sequences were as follows: ABB (132), ABB_ABS (54), ABB_ABS_FLD (45), ABB_ABS_PA14_FLD (20) and ABS_FLD_ABB (22). Multiple protein alignment was performed using ClustalW program, built into the software version 2.1. Phylogenetic analysis was conducted with the Ape package implemented for R programming language.

**Acknowledgements**- We thank the German Electron Synchrotron (Hamburg, Germany) for assistance at Petra III P13 Beamline, the Diamond Synchrotron Radiation Source (Daresbury, UK) for assistance at I03 Beamline, and funding from the European Community's Seventh Framework Programme under BioStruct-X (grant agreement N°283570). We also thank the staff of the European Synchrotron Radiation Facility at Grenoble (ESRF, France) for providing access and for technical assistance at beamlines ID23-1 and BM29, and the Spanish Synchrotron at Barcelona (ALBA, SPAIN) for assistance at XALOC beamline. We also acknowledge Rafael Bargiela for his excellent support in the preparation of Figures 1 and 5, and Oleg N Reva for his critical contribution to the compositional similarities analysis.

**Conflict of interest-** The authors declare that they have no conflicts of interest with the contents of this article.

**Author contributions**- JSA, MF and JP conceived and coordinated the study. MVP and MF contributed to screening, gene cloning and enzyme production and characterization. PNG produced metagenomic expression libraries and phylogeny analyses. JSA, BGP and MRE designed the crystallographic work and the SAXS experiments and interpreted the results. MRE performed all the crystallography and SAXS experiments. JMN and JP performed the phylogenetic analysis. JSA and MF wrote the paper, and all authors read and commented on the manuscript.

## REFERENCES


1. Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucl. Ac. Res.* **37**, D233-238
2. Lee, J. H., Hyun, Y. J., and Kim, D. H. (2011) Cloning and characterization of alpha-L-arabinofuranosidase and bifunctional alpha-L-arabinopyranosidase/beta-D-galactopyranosidase from *Bifidobacterium longum* H-1. *J. Appl. Microbiol.* **111**, 1097-1107
3. Mayer, C., Vocadlo, D. J., Mah, M., Rupitz, K., Stoll, D., Warren, R. A., and Withers, S. G. (2006) Characterization of a beta-N-acetylhexosaminidase and a beta-N-acetylglucosaminidase/beta-glucosidase from *Cellulomonas fimi*. *FEBS J.* **273**, 2929-2941
4. DeBoy, R. T., Mongodin, E. F., Fouts, D. E., Tailford, L. E., Khouri, H., Emerson, J. B., Mohamoud, Y., Watkins, K., Henrissat, B., Gilbert, H. J., and Nelson, K. E. (2008) Insights into plant cell wall degradation from the genome sequence of the soil bacterium *Cellvibrio japonicus*. *J. Bacteriol.* **190**, 5455-5463
5. Mai, V., Wiegel, J., and Lorenz, W. W. (2000) Cloning, sequencing, and characterization of the bifunctional xylosidase-arabinosidase from the anaerobic thermophile *Thermoanaerobacter ethanolicus*. *Gene* **247**, 137-143
6. Zhou, J., Bao, L., Chang, L., Liu, Z., You, C., and Lu, H. (2012) Beta-xylosidase activity of a GH3 glucosidase/xylosidase from yak rumen metagenome promotes the enzymatic degradation of hemicellulosic xylans. *Lett. Appl. Microbiol.* **54**, 79-87
7. Varghese, J. N., Hrmova, M., and Fincher, G. B. (1999) Three-dimensional structure of a barley beta-D-glucan exohydrolase, a family 3 glycosyl hydrolase. *Structure* **7**, 179-190
8. Pozzo, T., Pasten, J. L., Karlsson, E. N., and Logan, D. T. (2010) Structural and functional analyses of beta-glucosidase 3B from *Thermotoga neapolitana*: a thermostable three-domain representative of glycoside hydrolase 3. *J. Mol. Biol.* **397**, 724-739
9. Yoshida, E., Hidaka, M., Fushinobu, S., Koyanagi, T., Minami, H., Tamaki, H., Kitaoka, M., Katayama, T., and Kumagai, H. (2010) Role of a PA14 domain in determining substrate specificity of a glycoside hydrolase family 3 beta-glucosidase from *Kluyveromyces marxianus*. *Biochem. J.* **431**, 39-49
10. Nakatani, Y., Cutfield, S. M., Cowieson, N. P., and Cutfield, J. F. (2012) Structure and activity of exo-1,3/1,4-beta-glucanase from marine bacterium *Pseudoalteromonas sp*. BB1 showing a novel C-terminal domain. *FEBS J.* **279**, 464-478
11. Zmudka, M. W., Thoden, J. B., and Holden, H. M. (2013) The structure of DesR from *Streptomyces venezuelae*, a beta-glucosidase involved in macrolide activation. *Protein Sci.* **22**, 883-892
12. Karkehabadi, S., Helmich, K. E., Kaper, T., Hansson, H., Mikkelsen, N. E., Gudmundsson, M., Piens, K., Fujdala, M., Banerjee, G., Scott-Craig, J. S., Walton, J. D., Phillips, G. N., Jr., and Sandgren, M. (2014) Biochemical characterization and crystal structures of a fungal family 3 beta-glucosidase, Cel3A from *Hypocrea jecorina*. *J. Biol. Chem.* **289**, 31624-31637
13. Suzuki, K., Sumitani, J., Nam, Y. W., Nishimaki, T., Tani, S., Wakagi, T., Kawaguchi, T., and Fushinobu, S. (2013) Crystal structures of glycoside hydrolase family 3 beta-glucosidase 1 from *Aspergillus aculeatus*. *Biochem. J.* **452**, 211-221
14. Agirre, J., Ariza, A., Offen, W. A., Turkenburg, J. P., Roberts, S. M., McNicholas, S., Harris, P. V., McBrayer, B., Dohnalek, J., Cowtan, K. D., Davies, G. J., and Wilson, K. S. (2016) Three-dimensional structures of two heavily N-glycosylated *Aspergillus sp.* family GH3 beta-D-glucosidases. *Acta Cryst. D* **72**, 254-265
15. Marin-Navarro, J., Gurgu, L., Alamar, S., and Polaina, J. (2011) Structural and functional analysis of hybrid enzymes generated by domain shuffling between *Saccharomyces cerevisiae* (var. diastaticus) Sta1 glucoamylase and *Saccharomycopsis fibuligera* Bgl1 beta-glucosidase. *Appl. Microbiol. Biotechnol.* **89**, 121-130
16. Litzinger, S., Fischer, S., Polzer, P., Diederichs, K., Welte, W., and Mayer, C. (2010) Structural and kinetic analysis of *Bacillus subtilis* N-acetylglucosaminidase reveals a unique Asp-His dyad mechanism. *J. Biol. Chem.* **285**, 35675-35684

17. Bacik, J. P., Whitworth, G. E., Stubbs, K. A., Vocadlo, D. J., and Mark, B. L. (2012) Active site plasticity within the glycoside hydrolase NagZ underlies a dynamic mechanism of substrate distortion. *Chem. Biol.* **19**, 1471-1482
18. Nakajima, M., Yoshida, R., Miyanaga, A., Abe, K., Takahashi, Y., Sugimoto, N., Toyoizumi, H., Nakai, H., Kitaoka, M., and Taguchi, H. (2016) Functional and structural analysis of a beta-glucosidase involved in beta-1,2-glucan metabolism in *Listeria innocua. PLoS One* **11**, e0148870
19. Hrmova, M., De Gori, R., Smith, B. J., Fairweather, J. K., Driguez, H., Varghese, J. N., and Fincher, G. B. (2002) Structural basis for broad substrate specificity in higher plant beta-D-glucan glucohydrolases. *Plant Cell* **14**, 1033-1052
20. Hrmova, M., De Gori, R., Smith, B. J., Vasella, A., Varghese, J. N., and Fincher, G. B. (2004) Three-dimensional structure of the barley beta-D-glucan glucohydrolase in complex with a transition state mimic. *J. Biol. Chem.* **279**, 4970-4980
21. Hrmova, M., Streltsov, V. A., Smith, B. J., Vasella, A., Varghese, J. N., and Fincher, G. B. (2005) Structural rationale for low-nanomolar binding of transition state mimics to a family GH3 beta-D-glucan glucohydrolase from barley. *Biochemistry* **44**, 16529-16539
22. Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K. H., Whitman, W. B., Euzeby, J., Amann, R., and Rossello-Mora, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol*. **12**, 635-645
23. Alcaide, M., Tornes, J., Stogios, P. J., Xu, X., Gertler, C., Di Leo, R., Bargiela, R., Lafraya, A., Guazzaroni, M. E., Lopez-Cortes, N., Chernikova, T. N., Golyshina, O. V., Nechitaylo, T. Y., Plumeier, I., Pieper, D. H., Yakimov, M. M., Savchenko, A., Golyshin, P. N., and Ferrer, M. (2013) Single residues dictate the co-evolution of dual esterases: MCP hydrolases from the alpha/beta hydrolase family. *Biochem. J.* **454**, 157-166
24. Ferrer, M., Martinez-Martinez, M., Bargiela, R., Streit, W. R., Golyshina, O. V., and Golyshin, P. N. (2016) Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. *Microb. Biotechnol.* **9**, 22-34
25. Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., Toyohara, K., Miyamoto, K., Kimura, Y., and Oda, K. (2016) A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **351**, 1196-1199
26. Alcaide, M., Stogios, P. J., Lafraya, A., Tchigvintsev, A., Flick, R., Bargiela, R., Chernikova, T. N., Reva, O. N., Hai, T., Leggewie, C. C., Katzke, N., La Cono, V., Matesanz, R., Jebbar, M., Jaeger, K. E., Yakimov, M. M., Yakunin, A. F., Golyshin, P. N., Golyshina, O. V., Savchenko, A., Ferrer, M., and Consortium, M. (2015) Pressure adaptation is linked to thermal adaptation in salt-saturated marine habitats. *Environ. Microbiol.* **17**, 332-345
27. Gerlt, J. A., Allen, K. N., Almo, S. C., Armstrong, R. N., Babbitt, P. C., Cronan, J. E., Dunaway-Mariano, D., Imker, H. J., Jacobson, M. P., Minor, W., Poulter, C. D., Raushel, F. M., Sali, A., Shoichet, B. K., and Sweedler, J. V. (2011) The Enzyme Function Initiative. *Biochemistry* **50**, 9950-9962
28. Del Pozo, M. V., Fernandez-Arrojo, L., Gil-Martinez, J., Montesinos, A., Chernikova, T. N., Nechitaylo, T. Y., Waliszek, A., Tortajada, M., Rojas, A., Huws, S. A., Golyshina, O. V., Newbold, C. J., Polaina, J., Ferrer, M., and Golyshin, P. N. (2012) Microbial beta-glucosidases from cow rumen metagenome enhance the saccharification of lignocellulose in combination with commercial cellulase cocktail. *Biotechnol. Biofuels* **5**, 73
29. Menigaud, S., Mallet, L., Picord, G., Churlaud, C., Borrel, A., and Deschavanne, P. (2012) GOHTAM: a website for 'Genomic Origin of Horizontal Transfers, Alignment and Metagenomics'. *Bioinformatics* **28**, 1270-1271
30. Jami, E., Israel, A., Kotser, A., and Mizrahi, I. (2013) Exploring the bovine rumen bacterial community from birth to adulthood. *ISME J* **7**, 1069-1079
31. Pitta, D. W., Pinchak, W. E., Indugu, N., Vecchiarelli, B., Sinha, R., and Fulford, J. D. (2016) Metagenomic Analysis of the Rumen Microbiome of Steers with Wheat-Induced Frothy Bloat. *Front. Microbiol.* **7**, 689
32. McAndrew, R. P., Park, J. I., Heins, R. A., Reindl, W., Friedland, G. D., D'Haeseleer, P., Northen, T., Sale, K. L., Simmons, B. A., and Adams, P. D. (2013) From soil to structure, a novel dimeric beta-glucosidase belonging to glycoside hydrolase family 3 isolated from compost using metagenomic analysis. *J. Biol. Chem.* **288**, 14985-14992

33. Holm, L., and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D. *Nucl. Ac. Res.* **38**, W545-549
34. Hrmova, M., Varghese, J. N., De Gori, R., Smith, B. J., Driguez, H., and Fincher, G. B. (2001) Catalytic mechanisms and reaction intermediates along the hydrolytic pathway of a plant beta-D-glucan glucohydrolase. *Structure (London, England : 1993)* **9**, 1005-1016
35. Hegyi, H., and Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11**, 1632-1640
36. Bashton, M., and Chothia, C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.* **315**, 927-939
37. Ricard, G., McEwan, N. R., Dutilh, B. E., Jouany, J. P., Macheboeuf, D., Mitsumori, M., McIntosh, F. M., Michalowski, T., Nagamine, T., Nelson, N., Newbold, C. J., Nsabimana, E., Takenaka, A., Thomas, N. A., Ushida, K., Hackstein, J. H., and Huynen, M. A. (2006) Horizontal gene transfer from Bacteria to rumen Ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* **7**, 22
38. Berg Miller, M. E., Yeoman, C. J., Chia, N., Tringe, S. G., Angly, F. E., Edwards, R. A., Flint, H. J., Lamed, R., Bayer, E. A., and White, B. A. (2012) Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ. Microbiol.* **14**, 207-227
39. Hemsley, A., Arnheim, N., Toney, M. D., Cortopassi, G., and Galas, D. J. (1989) A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucl. Ac. Res.* **17**, 6545-6551
40. Kabsch, W. (2010) Xds. *Acta Cryst. D* **66**, 125-132
41. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011) Overview of the CCP4 suite and current developments. *Acta Cryst. D* **67**, 235-242
42. Vagin, A., and Teplyakov, A. (2010) Molecular replacement with MOLREP. *Acta Cryst. D* **66**, 22-25
43. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Cryst. D.* **53**, 240-255
44. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Cryst. D* **60**, 2126-2132
45. Kirschner, K. N., Yongye, A. B., Tschampel, S. M., Gonzalez-Outeirino, J., Daniels, C. R., Foley, B. L., and Woods, R. J. (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29**, 622-655
46. DeLano, W. L. (2002) Pymol: An open-source molecular graphics tool. *The PyMOL Molecular Graphics System, DeLano Scientific, San Carlos, CA, USA.*
47. Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D., Konarev, P. V., and Svergun, D. I. (2012) New developments in the program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342-350
48. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., and Schwede, T. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucl. Ac. Res.* **42**, W252-258
49. Nielsen, M., Lundegaard, C., Lund, O., and Petersen, T. N. (2010) CPHmodels-3.0--remote homology modeling using structure-guided sequence profiles. *Nucl. Ac. Res.* **38**, W576-581

**FOOTNOTES**

Research was supported by grants BIO2013-48779-C4-2-R, BIO2013-48779-C4-3-R, and BIO2014-54494-R from the Spanish Ministry of Economy and Competitiveness. Research was also funded by the ERA Net IB2 Project MetaCat through Spanish Ministry of Economy and Competitiveness, grant number PCIN-2014-107, and UK Biotechnology and Biological Sciences Research Council (BBSRC), grant BB/M029085/1. This project has received also funding from the European Union's Horizon 2020 research and innovation program [Blue Growth: Unlocking the potential of Seas and Oceans] under grant agreement No [634486]. The authors gratefully acknowledge the financial support provided by the European Regional Development Fund (ERDF).

The atomic coordinates and structure factors (codes 5K6L, 5K6M, 5K6N, 5K6O) have been deposited in the Protein Data Bank (http://wwpdb.org/).

The abbreviations used are: CBM, carbohydrate binding module; DP, degree of polymerization; GH, glycoside hydrolase; PEG, polyethyleneglycol; *p*NP, *p*-nitrophenyl; *p*NPαGlc, *p*NP-α-glucose; *p*NPαMal, *p*NP-α-maltose; *p*NPβGlc, *p*NP-β-glucose; *p*NPβCel, *p*NP-β-cellobiose; *p*NPαAraf, *p*NP-α-arabinofuranose; *p*NPβArap, *p*NP-β-arabinopyranose; *p*NPαXyl, *p*NP-α-xylose; *p*NPβXyl, *p*NP-β-xylose; *p*NPβXylb, *p*NP-β-xylobiose; *p*NPαFuc, *p*NP-α-fucose; *p*NPαRha, *p*NP-α-rhamnose; *p*NPαMan, *p*NP-α-mannose; *p*NPβMan, *p*NP-β-mannose; *p*NPαGal, *p*NP-α-galactose; *p*NPβGal, *p*NP-β-galactose; *p*NPβLac, *p*NP-β-lactose; *p*NPGlcNAc, *p*NP-N-acetyl-β-D-glucosaminide; *p*NPGalNAc, *p*NP-N-acetyl-β-D-galactosaminide; rmsd, root mean square deviation; SRF, Seminal rumen fluid.

**Table 1. Substrate specificity of the purified β-glycosidase $GlyA_1$ and truncated $GlyA_1$-ΔCt**

| | Specific activity (units/g) | |
|---|---|---|
| **Substrate** | **$GlyA_1$** | **$GlyA_1$-ΔCt** |
| *p*NPβGlc | 2226.8±120.2 | 197.0±10.6 |
| *p*NPβXyl | 2876.2±111.1 | 624.3±34.1 |
| *p*NPβXylb | 301.6±8.3 | 14.8±0.70 |
| *p*NPβCel | 290.7±9.9 | 17.04±0.50 |
| *p*NPβGal | 11.8±2.2 | 0.64±0.01 |
| *p*NPβFuc | 1.18±0.01 | 0.62±0.01 |
| *p*NPαAraf | 1.16±0.01 | 0.64±0.01 |
| *p*NPαArap | 0.66±0.01 | 0.33±0.01 |
| Cellobiose | 551.5±2.6 | 43.6±2.2 |
| Cellotriose | 532.6±0.5 | 44.4±4.9 |
| Cellotetraose | 569.3±0.4 | 48.7±5.1 |
| Cellopentaose | 641.5±0.5 | 53.3±2.1 |
| Xylobiose | 634.2±4.3 | 115.0±4.4 |
| Xylotriose | 668.3±7.2 | 136.1±6.7 |
| Xylotetraose | 674.5±11.5 | 174.7±4.5 |
| Xylopentaose | 747.9±3.5 | 196.4±8.6 |
| Gentibiose | 535.2±0.7 | 45.9±1.2 |
| Sophorose | 602.8±0.7 | 54.3±4.5 |
| Lichenan | 68.6±4.8 | 9.8±0.7 |

**Table 2 Kinetic parameters of the purified β-glycosidase $GlyA_1$**

| **Substrate** | **$K_M$ (mM)** | **$k_{cat}$ ($s^{-1}$)** | **$k_{cat}/K_M$ ($s^{-1}$ $M^{-1}$)** |
|---|---|---|---|
| *p*NPβGlc | 10.7±2.0 | 1.63±0.38 | 152.3 |
| *p*NPβXyl | 8.8±0.5 | 0.95±0.44 | 107.8 |
| *p*NPβCel | 1.4±0.2 | 0.51±0.15 | 314.2 |
| *p*NPβXylb | 2.5±0.3 | 0.73±0.12 | 292.0 |
| *p*NPβGal | 7.6±0.1 | 0.13±0.04 | 17.1 |
| *p*NPβFuc | 4.8±0.3 | 0.05±0.01 | 10.4 |
| *p*NPαAraf | 7.8±1.7 | 0.03±0.01 | 3.85 |
| *p*NPαArap | 10.7±3.4 | 0.01±0.01 | 0.93 |
| Celobiose | 2.4±0.3 | 0.07±0.01 | 28.2 |
| Xylobiose | 4.7±0.2 | 0.05±0.01 | 10.6 |

**Table 3. Crystallographic Data of $GlyA_1$**
(Values in parentheses are for the high resolution shell)

| **Crystal data** | $GlyA_1$/ Glycerol | $GlyA_1$/ Glucose | $GlyA_1$-ΔCt/ Xylose | $GlyA_1$-ΔCt/ Galactose |
|---|---|---|---|---|
| Space group | $P2_12_12_1$ | $P2_12_12_1$ | $P2_12_12_1$ | $P2_12_12_1$ |
| Unit cell parameters | | | | |
| a (Å) | 51.22 | 50.63 | 50.60 | 50.92 |
| b (Å) | 119.72 | 119.18 | 119.32 | 119.25 |
| c (Å) | 157.49 | 157.42 | 157.20 | 157.48 |
| **Data collection** | | | | |
| Beamline | Diamond (IO3) | PetraIII/ DESY (P13) | ESRF (ID23-1) | ALBA (XALOC) |
| Temperature (K) | 100 | 100 | 100 | 100 |
| Wavelength (Å) | 0.9762 | 0.9786 | 0.9762 | 1.1271 |
| Resolution (Å) | 95.31-1.83 (1.83-1.87) | 95.03-2.17 (2.17-2.24) | 95.05-2.08 (2.08-2.14) | 95.08-2.29 (2.29-2.37) |
| **Data processing** | | | | |
| Total reflections | 537914 (21607) | 338356 (28624) | 384429 (29703) | 287828 (27811) |
| Unique reflections | 84644 (3858) | 51199 (4369) | 58135 (4451) | 44188 (4264) |
| Multiplicity | 6.4 (5.6) | 6.6 (6.6) | 6.6 (6.7) | 6.5 (6.5) |
| Completeness (%) | 98.9 (87.2) | 99.7 (99.9) | 99.9 (99.9) | 100.0 (100.0) |
| Mean $I/\sigma$ ($I$) | 8.7 (2.1) | 11.0 (3.3) | 10.9 (3.0) | 11.4 (3.3) |
| $R_{merge}$ [†] (%) | 13.7 (56.2) | 12.7 (57.9) | 8.7 (54.0) | 9.0 (52.8) |
| $R_{pim}$ [††] (%) | 5.9 (24.9) | 5.4 (24.4) | 3.6 (22.4) | 3.8 (22.3) |
| Molecules / ASU | 1 | 1 | 1 | 1 |
| **Refinement** | | | | |
| $R_{work}$ / $R_{free}$ [†††] (%) | 15.68/17.85 | 17.48/21.82 | 17.38/21.25 | 18.16/22.72 |
| **Nº of atoms/ average B** (Å²) | | | | |
| Protein | 6150 /19.55 | 6121/ 31.57 | 6129 / 41.37 | 6151 / 47.72 |
| Carbohydrate | 0 / 0 | 12 / 38.94 | 10 / 43.17 | 12 / 53.10 |
| Water molecules | 674 / 28.82 | 328 / 31.23 | 304 / 41.10 | 120 / 39.00 |
| All atoms | 6878 /20.66 | 6466 / 31.59 | 6503 / 41.66 | 6331 / 47.88 |
| **Ramachandran plot** (%) | | | | |
| Favoured | 98.00 | 97.00 | 98.00 | 98.00 |
| Outliers | 0 | 0 | 0 | 0 |
| **RMS deviations** | | | | |
| Bonds (Å) | 0.007 | 0.007 | 0.008 | 0.010 |
| Angles (°) | 1.209 | 1.218 | 1.259 | 1.417 |
| **PDB codes** | 5K6L | 5K6M | 5K6N | 5K6O |

[†] $R_{merge} = \Sigma_{hkl} \Sigma_i | I_i(hkl) - [I(hkl)] | / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the ith measurement of reflection hkl and [I(hkl)] is the weighted mean of all measurements.

[††] $R_{pim} = \Sigma_{hkl} [1/(N - 1)] 1/2 \Sigma_i | I_i(hkl) - [I(hkl)] | / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where N is the redundancy for the hkl reflection.

[†††] $R_{work} / R_{free} = \Sigma_{hkl} | Fo - Fc | / \Sigma_{hkl} | Fo |$, where Fc is the calculated and Fo is the observed structure factor amplitude of r eflection hkl for the working/free (5%) set.

**Table 4. SAXS Data Collection and derived parameters**

| Protein | Merged Data (mg/ml) | From Guinier | | | From Gnom | | | From Porod | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rg (nm) | Quality | I0 | Rg (nm) | I0 | Dmax | Porod Volumen ($nm^3$) | Mw (KDa) |
| $GlyA_1$-ΔCt | 0.34-5.29 | 2.784 | 88% | 76.38 | 2.80 | 76.04 | 8.479 | 122.71 | 72.18 |
| $GlyA_1$ | 0.32-5.04 | 2.918 | 84% | 83.41 | 2.95 | 83.92 | 9.120 | 148.76 | 87.51 |

## FIGURE LEGENDS

**Figure 1. Temperature (A) and pH (B) profiles of the purified β-glucosidase $GlyA_1$.** The data represent the relative percentages of specific activity (units/g) compared with the maximum activity using *p*NPβGlc as substrate (100% in panel A: 2841 units/g; 100% in panel B: 3056 units/g). The specific activities were calculated using 0.23 μM of protein and 10 mg/ml *p*NPβGlc as the assay substrate. For Panel A, reactions were performed in 50 mM sodium acetate buffer, pH 5.6, at different temperatures. For Panel B, reactions were performed at different pH (50 mM BR buffer) and 40ºC. Standard deviations of the results of assays conducted in triplicate are shown.

**Figure 2. The permuted domain composition of $GlyA_1$.** A) Comparison of $GlyA_1$ structure with representative members of multidomain GH3 enzymes: the β-glucosidases from *K. marxianus*, *Km*βGlu (9) and *T. neapolitana*, *Tn*βGlu (8), the exo-1,3/1,4-β-glucanase from *Pseudoalteromonas* sp., *Ps*ExoP (10) and the barley β-D-glucan exohydrolase, *Hv*ExoI (7). Domains are named as ABS: $(\alpha/\beta)_6$-sandwich; FLD fibronectin-like; ABB $(\alpha/\beta)_8$ barrel; PA14, protective antigen PA14 domain. B) Folding of $GlyA_1$: the N-terminal $(\alpha/\beta)_6$-sandwich domain (red) is followed by the FnIII domain (beige) and the $(\alpha/\beta)_8$ barrel domain (green). Two long segments connect the three domains (grey). A glucose found in the active site is represented in spheres. C) Scheme of the $GlyA_1$ domain organisation (left) as compared to that of *T. neapolitana* β-glucosidase (right) (8). D) Superimposition of $GlyA_1$ (gold) onto *T. neapolitana* β-glucosidase (blue) coordinates. Both enzymes present a deviation from the canonical $(\alpha/\beta)_8$ barrel topology, with their first α-helix missing, which makes strand β2 reversed and antiparallel with the other seven strands. The main difference between both enzymes is the long arm linking the FnIII to the $(\alpha/\beta)_8$ domain in $GlyA_1$, which is missing in *T. neapolitana* β-glucosidase. Also, small differences in the orientation of some helixes are observed.

**Figure 3. $GlyA_1$ active site architecture.** A) Detail of the loops surrounding its active site from the $(\alpha/\beta)_8$ barrel (green) and the $(\alpha/\beta)_6$-sandwich (raspberry) domains, superimposed onto the *T. neapolitana* β-glucosidase (8) (pale blue). Three glycerol molecules from the cryobuffer found in the $GlyA_1$ crystals are shown in orange. Asp709 and Glu143 are the nucleophile and the acid/base catalyst, respectively. Main features of $GlyA_1$ are the extended loop containing Asp709, which includes Trp711 and the ion-pair Arg717-Glu447 fixing it to the unique long arm, and a highly flexible loop containing Trp106. Two different conformations found among the crystals at Trp111 and Phe147 are highlighted. B) Detail of the atomic interactions defining subsite -1; a glucose molecule is shown in gold. Xylose binds in the same relaxed chair conformation and only interaction of the glucose O6 hydroxyl is missing. Inset: binding mode of galactose in a semi-chair conformation by flattening of the C4 atom that has the axial hydroxyl substituent and keeping the same interactions pattern. C) thiocellobiose (cyan) and thiogentibiose (pink) modelled at the active site by structural superimposition to the previously determined β-D-glucan glucohydrolase barley complexes (PDB entries 1IEX and 3WLP (34) ), delineating putative subsite +1. D) Molecular surface of the $GlyA_1$ active site, with relevant residues as sticks. Four different β−1,4/β-1,3−

linked tetraglucosides have been manually docked by superposition of their non-reduced end to the experimental glucose: a cellotetraose, as found in PDB entry 2Z1S (green), a Glc-4Glc-3Glc-4Glc (purple) and a Glc-4Glc-4Glc-3Glc (yellow), as built by the online carbohydrate-building program GLYCAM (45) and exported in its minimum energy state. E) Superposition of $GlyA_1$-Glc structure (beige) with those reported for *T. reesei* β-glucosidase (12) (purple) and barley β-D-glucan glucohydrolase complexed with thiocellobiose (34) (cyan).

**Figure 4. SAXS analysis of $GlyA_1$.** Six *ab initio* models were generated for complete $GlyA_1$ from SAXS data, using the experimental structure of the truncated protein and two different models of the last 120 residues ($GlyA_1$-Ct). The two templates were obtained from Swiss-Model (48) (in red colours) or CPH-model (49) (blue colours) servers, which predict different length of the linker attaching this domain to the core protein, 32 or 5 residues, respectively. CORAL (47) modelling of this linker in each run is represented in spheres. The active site pocket is indicated by the galactose found at the crystal (yellow) and the mobile loop (residues 100-113), as observed in the galactose-soaked crystals, is highlighted in green.

**Figure 5. $GlyA_1$ phylogenetic analysis.** The unrooted circular Neighbor-Joining tree indicating phylogenetic positions of polypeptide sequences of the $GlyA_1$ enzyme characterized in present work (in red boldface) and reference similar enzymes. GenBank or PDB (in boldface) accession numbers are indicated. The domain architecture (ABB, ABB_ABS, ABB_ABS_FLD, ABB-ABS(PA14)-FLD and ABS_FLD_ABB) to which each sequence associated is specifically indicated. Multiple protein alignment was performed using ClustalW program, built into the software version 2.1. Phylogenetic analysis was conducted with the Ape package implemented for R programming language. Sequences resembling NagZ (β-N-acetyl-glucosaminidase) are highlighted with pink background color. Those encoding GH3 β-glucosidases are indicated in brown color; within them, those with $GlyA_1$-like permuted domain topology are indicated in grey color. ABB, $(\alpha/\beta)_8$ barrel; ABS $(\alpha/\beta)_6$-sandwich; FLD fibronectin-like type III domain; PA14, protective antigen PA14 domain.

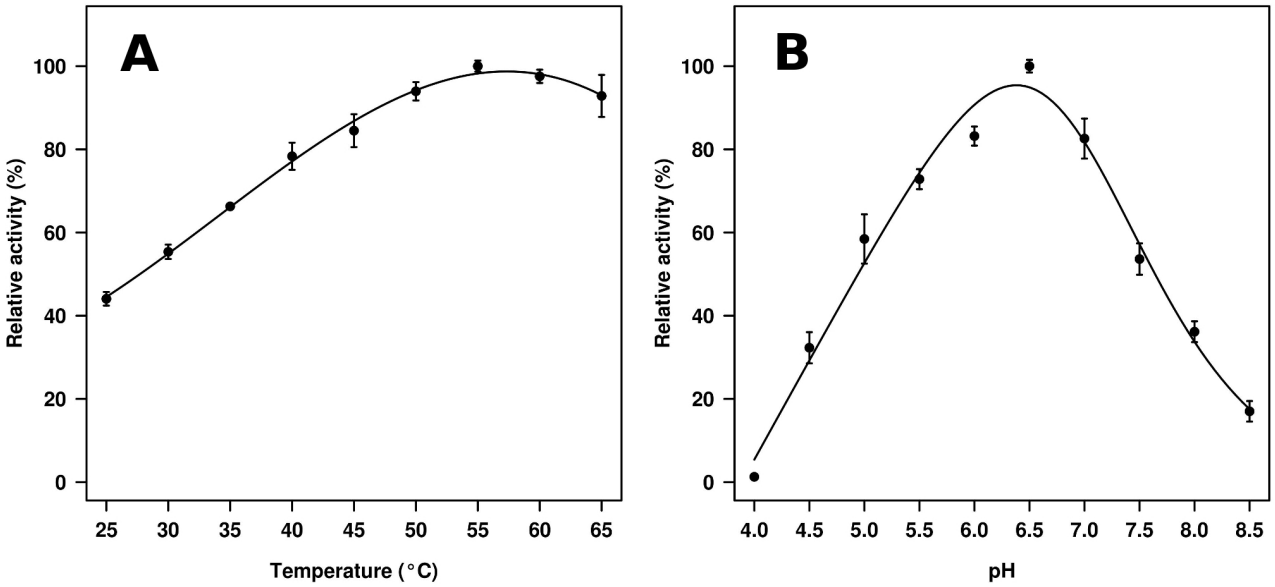

A
Relative activity (%)
0
20
40
60
80
100
25
30
35
40
45
50
55
60
65
Temperature (°C)
B
Relative activity (%)
0
20
40
60
80
100
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
pH

Fig 1

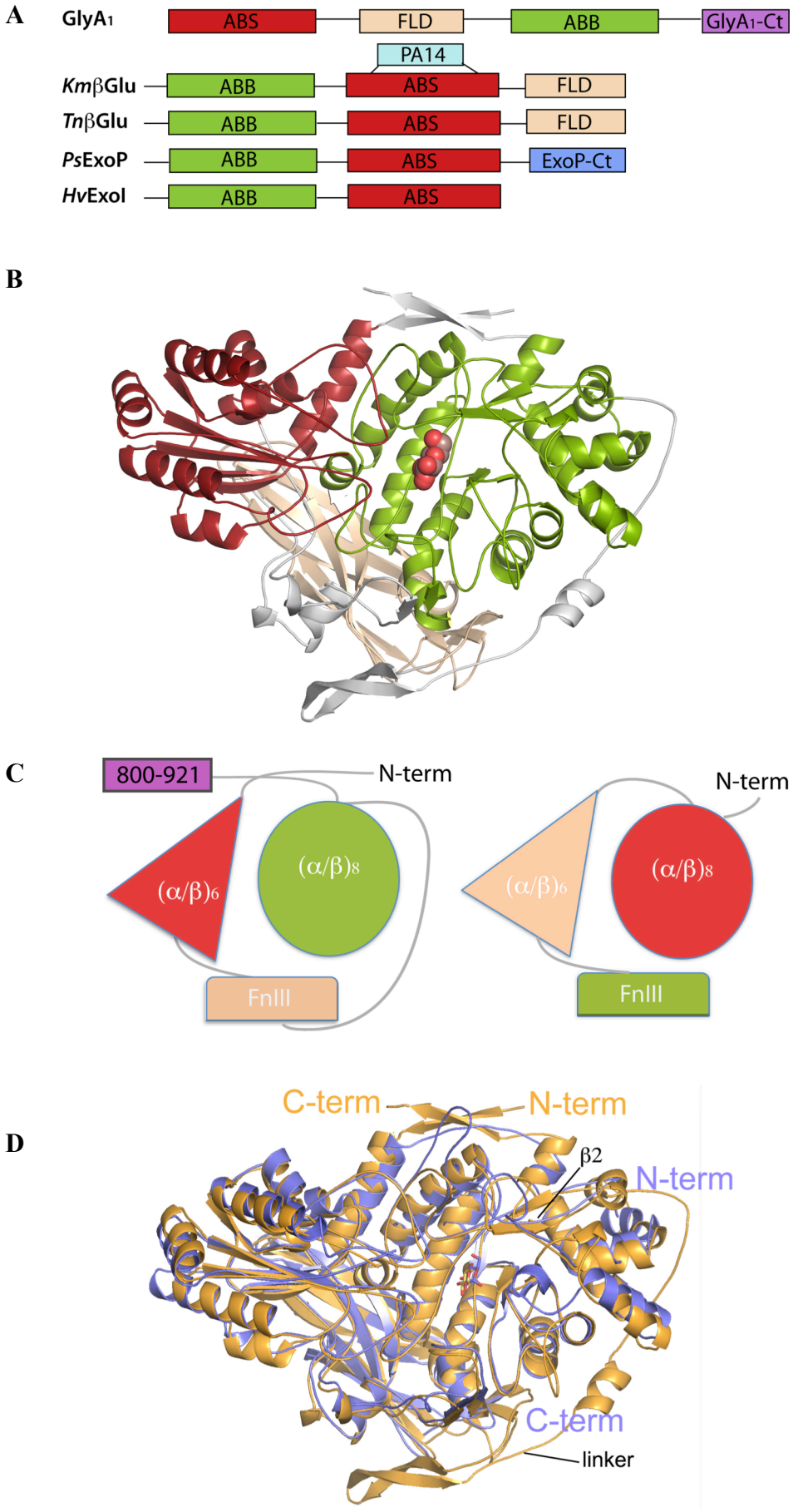
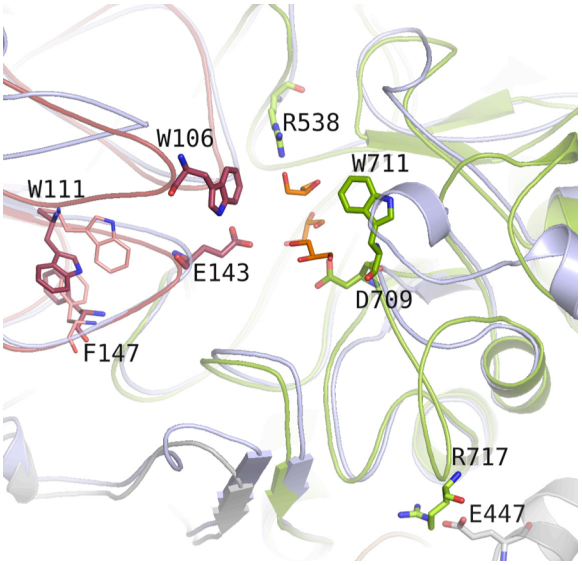
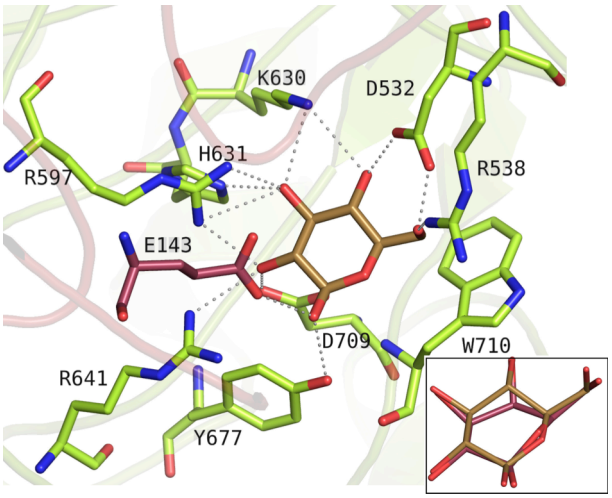

A
GlyA1
ABS
FLD
ABB
GlyA1-Ct
PA14
KmβGlu
ABB
ABS
FLD
TnβGlu
ABB
ABS
FLD
PsExoP
ABB
ABS
ExoP-Ct
HvExoI
ABB
ABS
B
C
800-921
N-term
(α/β)6
(α/β)8
FnIII
N-term
(α/β)6
(α/β)8
FnIII
D
C-term
N-term
β2
N-term
C-term
linker

Fig 2

**A**


W106
R538
W711
W111
E143
D709
F147
R717
E447

**B**


K630
D532
H631
R597
R538
E143
D709
W710
R641
Y677

**C**

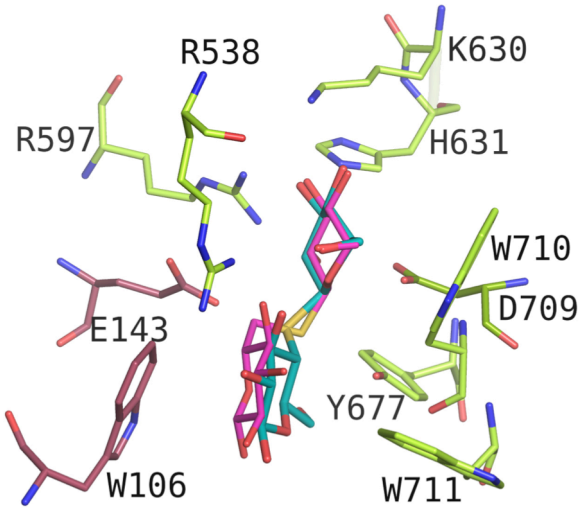
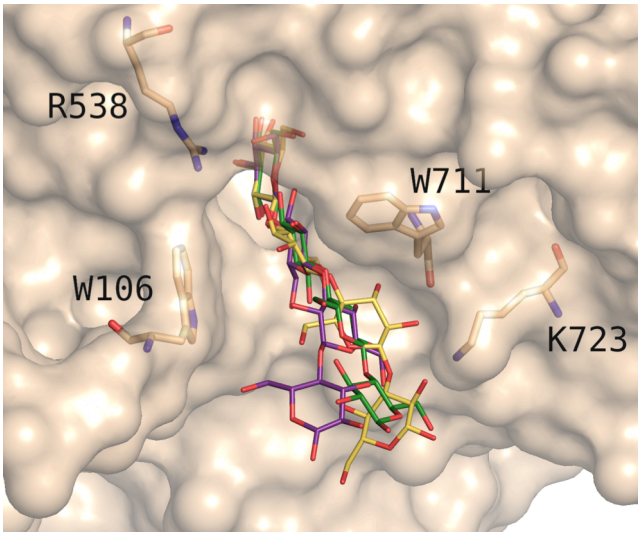

R538
K630
R597
H631
W710
D709
E143
Y677
W106
W711

Fig 3

**D**




**E**

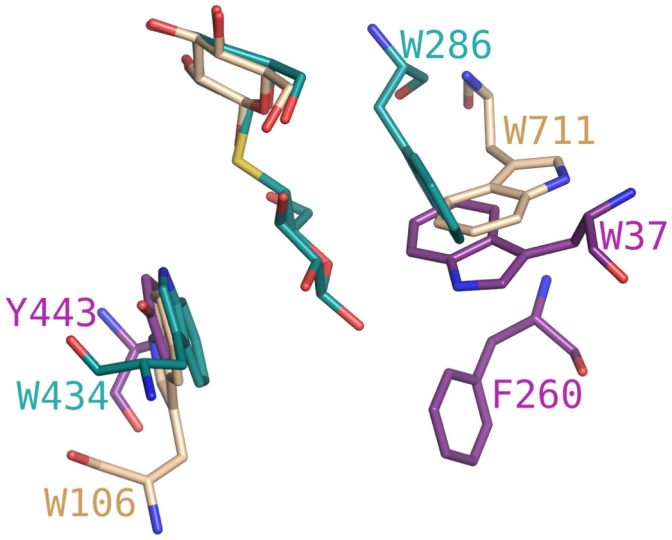



Fig 3 (cont)

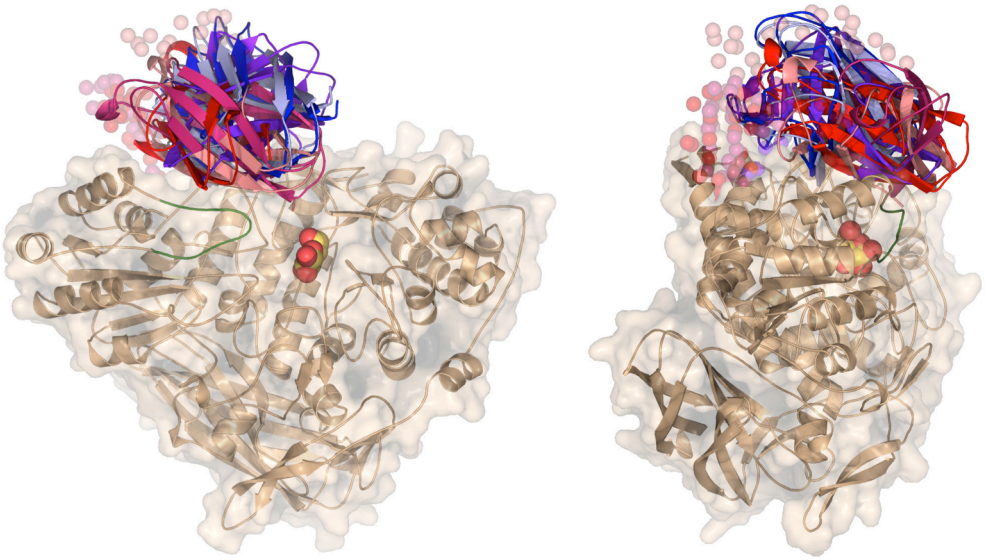


Fig 4

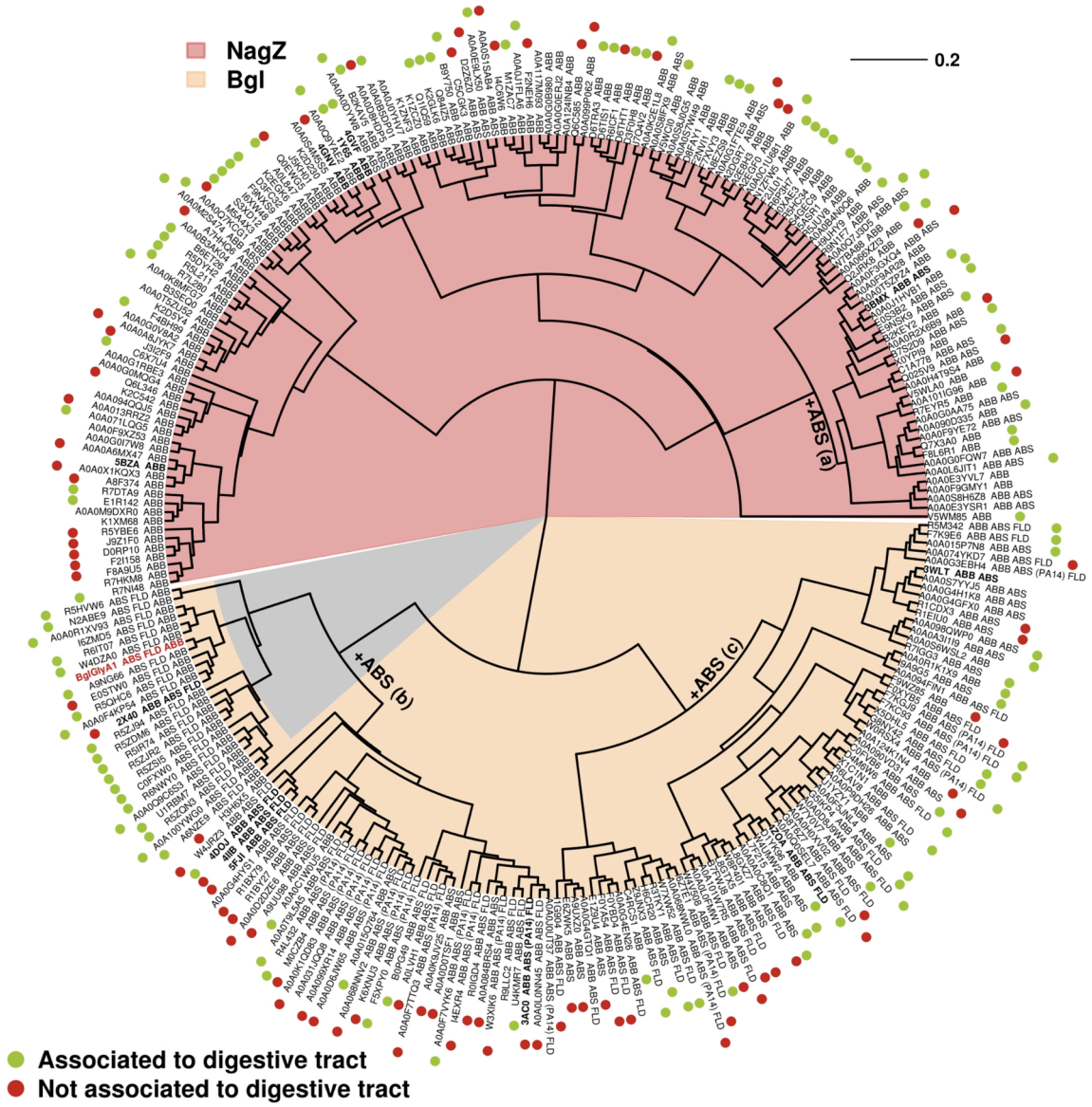

NagZ
Bgl
0.2
+ABS (a)
+ABS (b)
+ABS (c)
Associated to digestive tract
Not associated to digestive tract

Fig 5