# A subfamily classification to choreograph the diverse activities within glycoside hydrolase family 31

Thimali Arumapperuma, Jinling Li, Bastian Hornung, Niccolay Madiedo Soler, Ethan Goddard-Borger, Nicolas Terrapon, Spencer Williams

JBC RESEARCH ARTICLE

# A subfamily classification to choreograph the diverse activities within glycoside hydrolase family 31

Received for publication, December 13, 2022, and in revised form, February 1, 2023 Published, Papers in Press, February 17, 2023,
https://doi.org/10.1016/j.jbc.2023.103038

**Thimali Arumapperuma[1], Jinling Li[1], Bastian Hornung[2], Niccolay Madiedo Soler[3,4], Ethan D. Goddard-Borger[3,4], Nicolas Terrapon[2], and Spencer J. Williams[1,*]**

*From the [1]School of Chemistry and Bio21 Molecular Science and Biotechnology Institute and University of Melbourne, Parkville, Victoria, Australia; [2]AFMB, UMR 7257 CNRS Aix-Marseille Univ., USC 1408 INRAE, Marseille, France; [3]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia; [4]Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia*

Reviewed by members of the JBC Editorial Board. Edited by Chris Whitfield

**The Carbohydrate-Active Enzyme classification groups enzymes that breakdown, assemble, or decorate glycans into protein families based on sequence similarity. The glycoside hydrolases (GH) are arranged into over 170 enzyme families, with some being very large and exhibiting distinct activities/specificities towards diverse substrates. Family GH31 is a large family that contains more than 20,000 sequences with a wide taxonomic diversity. Less than 1% of GH31 members are biochemically characterized and exhibit many different activities that include glycosidases, lyases, and transglycosidases. This diversity of activities limits our ability to predict the activities and roles of GH31 family members in their host organism and our ability to exploit these enzymes for practical purposes. Here, we established a subfamily classification using sequence similarity networks that was further validated by a structural analysis. While sequence similarity networks provide a sequence-based separation, we obtained good segregation between activities among the subfamilies. Our subclassification consists of 20 subfamilies with sixteen subfamilies containing at least one characterized member and eleven subfamilies that are monofunctional based on the available data. We also report the biochemical characterization of a member of the large subfamily 2 (GH31_2) that lacked any characterized members: *Ra*GH31 from *Rhodoferax aquaticus* is an α-glucosidase with activity on a range of disaccharides including sucrose, trehalose, maltose, and nigerose. Our subclassification provides improved predictive power for the vast majority of uncharacterized proteins in family GH31 and highlights the remaining sequence space that remains to be functionally explored.**

Glycoside hydrolases (GHs) are present in all domains of life, and in viruses, and are involved in processes such as nutrient acquisition for bioenergetic metabolism, cell wall remodeling, glycoprotein biosynthesis and degradation, and pathogenesis. The Carbohydrate Active enZyme (CAZy; www.cazy.org; see also www.cazypedia.org) classification currently reports more than 170 GH families, with approximately five novel families released per year during the last decade (1, 2). The creation of a new family requires the characterization of at least one founding member and involves gathering the largest diversity of homologous proteins based on high sequence similarity. Later, based on literature survey and direct communications, the CAZy curators update the known family activities with more characterized members, which might either support its specificity or increase its diversity. Due to the ever-increasing rates at which genomes and metagenomes are sequenced, the growth of families far outpaces the ability of the research community to conduct experimental characterization, meaning that most members within each family are, and will remain, uncharacterized. For large families with diverse activities, this can limit our ability to understand and predict the roles of family members in their original host organism and our opportunities to exploit enzymes for practical purposes.

A solution to functional annotation of the dizzying deluge of sequence data lies in the creation of a subfamily-level classification that divides the family into smaller groups that might display more specific activities. Subclassifications can enhance the predictive power of sequence-based annotations, assisting in assigning likely activities to the uncharacterized members, and can guide the identification of unexplored regions of sequence space to be targeted for future exploration. The CAZy database implemented subfamily classifications based on phylogenies for families GH5 (3), GH13 (4), GH30 (5), and many polysaccharide lyases (6–9) or on Sequence Similarity Networks (SSNs) for families GH16 (10) and GH43 (11). The SSN method allows the analysis of very large datasets that could not reliably produce multiple sequence alignments and phylogenies.

The GH family 31 (GH31) is a large family with more than 20,000 sequences (by December 2022) from GenBank reported on CAZy website (www.cazy.org). Among these, only 130 proteins have been characterized, which exhibit activity on a variety of α-glycoside substrates and which have been linked to fifteen distinct enzyme commission (EC) numbers (Table 1). One of the most common enzyme activities in family GH31 is α-glucosidase (EC 3.2.1.20), which is defined by the EC as

* For correspondence: Spencer J. Williams, sjwill@unimelb.edu.au.

**Table 1**
**Fifteen EC numbers and corresponding enzyme activities reported to date within GH31 family**

| EC number | Activity |
|---|---|
| 3.2.1.10 | oligo-1,6-α-glucosidase |
| 3.2.1.11 | dextranase |
| 3.2.1.20 | α-glucosidase |
| 3.2.1.22 | α-galactosidase |
| 3.2.1.24 | α-mannosidase |
| 3.2.1.48 | sucrose α-glucosidase |
| 3.2.1.84 | glucan 1,3-α-glucosidase |
| 3.2.1.177 | α-D-xyloside xylohydrolase |
| 3.2.1.199 | sulfoquinovosidase |
| 3.2.1.204 | 1,3-α-isomaltosidase |
| 3.2.1.217 | exo-acting α-*N*-acetylgalactosaminidase |
| 2.4.1.24 | 1,4-α-glucan 6-α-glucosyltransferase |
| 2.4.1.161 | oligosaccharide 4-α-D-glucosyltransferase |
| 2.4.1.387 | isomaltosyltransferase |
| 4.2.2.13 | exo-(1→4)-α-D-glucan lyase |

EC number 3.2.1.28 corresponding to α,α-trehalase, is report in this study.

comprising enzymes whose specificity is mainly directly to the exohydrolysis of α-1,4-glucosidic linkages on maltooligosaccharides to liberate α-glucose. As is sometimes observed in other families, some GH31 enzymes exhibit a secondary, lower level of activity. For example, some α-glucosidases also exhibit oligosaccharide α-1,6-glucohydrolase (3.2.1.10) (12), α-mannosidase (3.2.1.24) (13), α-xylosidase (3.2.1.177) (14), and various transglucosidase (2.4.1.-) activities (15). Other examples include *Tropaeolum majus* α-xylosidase (3.2.1.177), which is also active on a range of α-glucosides, liberating glucose (16). Additional notable enzyme activities within GH31 family include sulfoquinovosidases (3.2.1.199) (17), α-glucan lyases (4.2.2.13) (18), and α-*N*-acetylgalactosaminidases (3.2.1.217) (19).

Despite the diversity of members and activities, enzymes within family GH31 use just three closely related mechanisms that all involve two strategically positioned carboxyl residues within their active site. GHs within this family operate with a stereochemically retaining mechanism, converting an α-glycoside substrate into the α-configured sugar hemiacetal. These enzymes use a classical Koshland two-step, double displacement mechanism, *via* a glycosyl-enzyme intermediate, with the two carboxyl residues acting as nucleophile and general acid/base (Fig. 1*A*) (20). In the first step, a carboxylate residue acts as nucleophile to form a glycosyl-enzyme intermediate, while a carboxylic acid residue acts as general acid to assist the departure of the anomeric group. In the second step, a water molecule performs a nucleophilic substitution at the anomeric center of the glycosyl-enzyme intermediate with the second residue now a carboxylate acting as general base. In the case of transglycosidases, the first step is identical, while in the second step, the nucleophilic water is replaced by an alcohol, most commonly the hydroxyl group of another sugar, resulting in the synthesis of a new glycoside (Fig. 1*B*). Finally, the α-glucan lyases of family GH31 catalyze the cleavage of α-1,4-glucosidic linkages in starch, glycogen, and maltooligosaccharides to form 1,5-anhydro-D-fructose. Kinetic isotope effects, linear free energy relationships, and intermediate trapping studies indicate that the α-glucan lyase mechanism involves two steps that share similarity to the double-displacement hydrolase mechanism (21). The first step yields a covalent glycosyl-enzyme intermediate in the same way as for GHs and transglycosidases. However, in the second step, a syn-elimination of the glycosyl-enzyme intermediate results in the formation of an unsaturated enol that can tautomerize to 1,5-anhydro-D-fructose (Fig. 1*C*).

Given the diversity of enzymatic activities in GH31 and that more than 99% GH31 members remain uncharacterized, we conducted a detailed subfamily classification for enzymes within family GH31, starting with SSN analysis, and complemented with structural analysis. Our subclassification includes 20 subfamilies, 15 of which already include a characterized member and cover 80% of the family members. Eleven of these subfamilies have a unique EC number, and three subfamilies have two, improving the predictive power and quality of GH31 annotations. Six subfamilies did not contain any characterized member, indicating to the research community possible regions of the sequence space that remains to be functionally explored. We characterized a member of the largest of these unexplored subfamilies, subfamily GH_2, moving from 80% to 96% the percentage of GH31 in a characterized subfamily. We show that *Ra*GH31 from *Rhodoferax aquaticus* is an α-glucosidase with a preference for sucrose and with significant activity on nigerose, trehalose, and maltose.

## Results

### Subfamily delineation

More than 13,000 GH31 modules, extracted from CAZy in June 2021, were subjected to pairwise all-*versus*-all BLAST alignments, as the first step for SSN analysis. Decreasing BLAST *E*-value thresholds ranging from $10^{-60}$ to $10^{-140}$ by steps of $10^{-5}$, hereafter referred as to the SSN *E*-value, or simply *E*, allowed the construction of a series of 17 SSNs. These 17 SSNs represent many options for the division of family GH31 into an increasing number of subfamilies (expected to increase in functional specificity) and of unclassified sequences (Table 2). Our objective was to identify the optimal SSN *E*-value, and corresponding subclassification scheme, that would provide the best trade-off in maximizing (i) the number of members classified into a subfamily, (ii) the robustness of automatic annotation (hidden Markov models, HMMs), and (iii) the utility from a functional-predictive perspective.

We first examined the distribution of EC numbers across subfamilies. At $E = 10^{-60}$, the largest subfamily gathers 95% of the GH31 members and most EC numbers (14/15), while EC 3.2.1.217 uniquely resides in one of two other small subfamilies. The third and last subfamily contains EC 3.2.1.22 (α-galactosidase), which is also found in the large first subfamily. By $E = 10^{-80}$, the largest group gave birth to four additional subfamilies, one fairly large (13% of the family) is specific to EC 3.2.1.199 (sulfoquinovosidase), one fairly small that contains EC 3.2.1.22 (α-galactosidase) and which is the last with this activity to separate, while the two others display activities also found in the largest subfamily, which still gathers 76% of GH31 members. Hence, most activities remain together

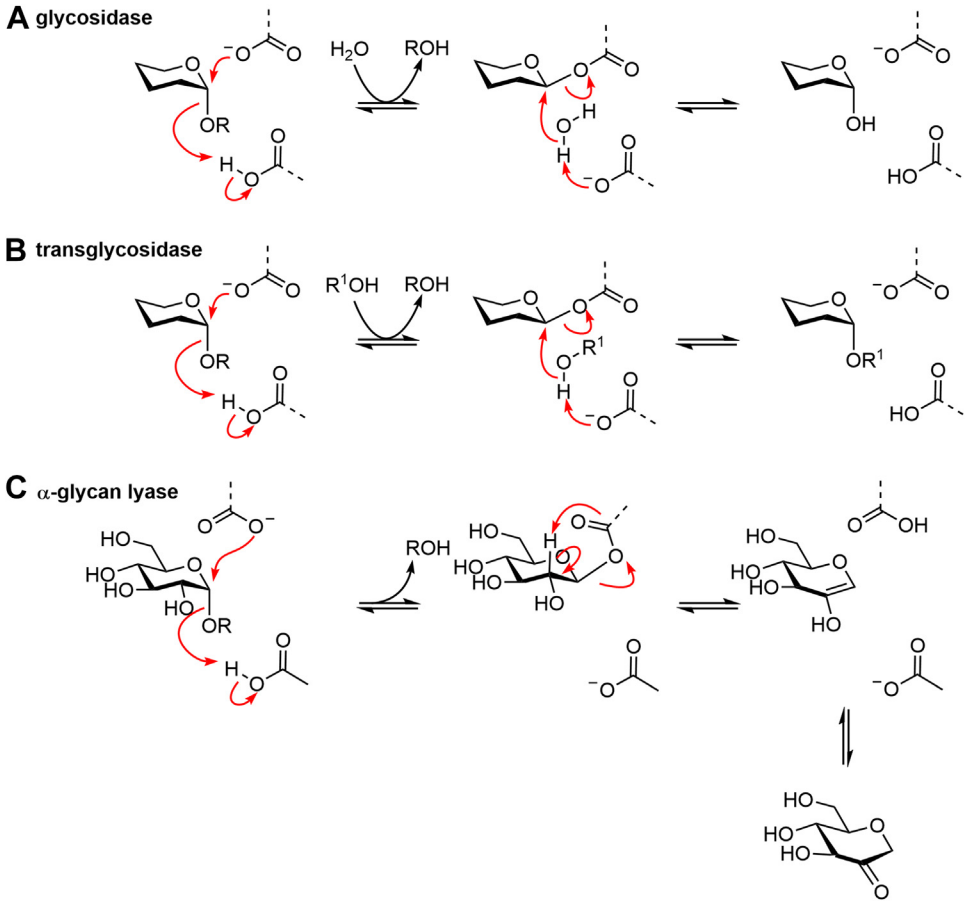


**Figure 1. Mechanisms for GH31 enzymes.** *A*, mechanism for retaining α-glycoside hydrolases proceeding through a glycosyl-enzyme intermediate. Sugar substituents have been omitted for clarity. *B*, mechanism for retaining glycosyl transfer catalyzed by transglucosidases. *C*, reaction catalyzed by α-glucan lyase involving *syn*-elimination of the glycosyl enzyme intermediate. GH, glycoside hydrolase.

in the largest subfamily (12/15). EC 4.2.2.13 (α-glucan lyase), the only lyase activity reported in GH31,[22] and EC 3.2.1.204 (α-isomaltosidase) along with associated EC 2.4.1.387 (3-α-isomaltosyltransferase) separate into specific subfamilies by $10^{-110}$, while the largest subfamily still gathers 75% of the GH31 members. By $E = 10^{-125}$, the largest subfamily further splits into several subfamilies more comparable in size: (i) the vestiges of the large family still gathering nine distinct EC numbers but now representing less than 16% of the family (∼2100 members); (ii) an equal-size subfamily which lacked any characterized members (but is investigated later in this article); (iii) a larger subfamily (>3500 members; 27% of GH31) with only two assigned EC numbers, 3.2.1.177 and 3.2.1.20; and (iv) three subfamilies (with 1594, 4, and 333 members; representing 14% of family GH31) that include two assigned the EC number 3.2.1.177 and one lacking an EC number. At even higher *E*-values, several additional subgroups emerged from these three large subfamilies, but these were small, with several distinguished by only taxonomy (Table S2), while some shared EC numbers with their sibling subfamilies, suggesting that the threshold is too stringent.

While consideration of EC number segregation is an essential qualitative criterion for a useful subclassification scheme, it also helps to guide the selection of the appropriate SSN *E*-value threshold by controlling the performance of the analysis through a more quantitative approach, notably by the ability of bioinformatics tools to correctly predict subfamily membership (10). Predictions were produced from HMM subfamily libraries for most SSN *E*-values. Overall precision and recall suggested that the most performant SSN *E*-values span $10^{-115}$, $10^{-125}$, and $10^{-130}$ thresholds, for which high precision and recall confirm the satisfactory EC number distribution previously discussed (Table S1). Finally, regarding the number of GH31 sequences that could not be assigned to any subfamily, they represent only a very low fraction of the total, frequently increasing by <5 members at each $10^{-5}$ SSN step, rarely getting close to 20, until the last two steps ($10^{-135}$ and $10^{-140}$), which each increased by >20 the nonclassified GH31 sequences, further supporting the previously preferred threshold. Altogether, we considered $10^{-125}$ the most appropriate SSN *E*-value for GH31 functional subclassification, and a visualization of the corresponding 20 subfamilies was generated using Cytoscape (Fig. 2). Analogous to previous GH subfamily classifications (10), the GH31 subfamilies were systematically referenced as "GH31_*n*," where *n* is the subfamily number.

Based on this subclassification, we performed a structural comparison of subfamilies by selecting an experimentally

**Table 2**
**GH31 subfamily analysis based on Sequence Similarity Networks constructed for decreasing BLAST *E*-value thresholds**

| Clan D, GH31 - 13 473 modules | *E* value | | | | | | | | | | | | | | | | | GH31_*n* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 | 100 | 105 | 110 | 115 | 120 | 125 | 130 | 135 | 140 | |
| | 12737 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.22 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.199 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 12118 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.22 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.199 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 12048 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.199 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 12044 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.199 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 10218 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 10211 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 10080 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 10080 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 10068 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 10038 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 3.2.1.204 2.4.1.24 2.4.1.161 2.4.1.387 4.2.2.13 | 9739 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 2.4.1.24 2.4.1.161 | 7806 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 2.4.1.24 2.4.1.161 | 4231 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 2.4.1.24 2.4.1.161 | 2092 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.48 3.2.1.84 3.2.1.177 2.4.1.24 2.4.1.161 | 1830 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.84 2.4.1.24 2.4.1.161 | 1689 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.84 2.4.1.24 | 1669 3.2.1.10 3.2.1.11 3.2.1.20 3.2.1.24 3.2.1.84 2.4.1.24 | 1 |
| | | | | | | | | | | | | | | | 141 2.4.1.161 | 141 2.4.1.161 | 141 2.4.1.161 | |
| | | | | | | | | | | | | | | | 262 3.2.1.10 3.2.1.20 3.2.1.48 3.2.1.84 3.2.1.177 | 262 3.2.1.10 3.2.1.20 3.2.1.48 3.2.1.84 3.2.1.177 | 262 3.2.1.10 3.2.1.20 3.2.1.48 3.2.1.84 3.2.1.177 | |
| | | | | | | | | | | | | | | 2138 | 2102 | 2102 | 2102 | 2 |
| | | | | | | | | | | | | | | | 36 | 36 | 36 | |
| | | | | | | | | | | | | | 3573 3.2.1.20 3.2.1.177 | 3573 3.2.1.20 3.2.1.177 | 3507 3.2.1.20 3.2.1.177 | 3507 3.2.1.20 3.2.1.177 | 3507 3.2.1.20 3.2.1.177 | 3 |
| | | | | | | | | | | | | | | | 63 | 54 | 54 | |
| | | | | | | | | | | | | 1598 3.2.1.177 | 1598 3.2.1.177 | 1594 3.2.1.177 | 1594 3.2.1.177 | 1594 3.2.1.177 | 1594 3.2.1.177 | 4 |
| | | | | | | | | | | | | | | 4 3.2.1.177 | 4 3.2.1.177 | 4 3.2.1.177 | 4 3.2.1.177 | 5 |
| | | | | | | | | | | | | 333 | 333 | 333 | 333 | 333 | 333 | 6 |
| | | | | | | | | | | | 267 2.4.1.387 3.2.1.204 | 267 2.4.1.387 3.2.1.204 | 267 2.4.1.387 3.2.1.204 | 267 2.4.1.387 3.2.1.204 | 267 2.4.1.387 3.2.1.204 | 267 2.4.1.387 3.2.1.204 | 267 2.4.1.387 3.2.1.204 | 7 |
| | | | | | | | | | | | 24 4.2.2.13 | 24 4.2.2.13 | 24 4.2.2.13 | 24 4.2.2.13 | 24 4.2.2.13 | 24 4.2.2.13 | 24 4.2.2.13 | 8 |
| | | | | | | | | | | 26 | 26 | 26 | 26 | 26 | 26 | 26 | 26 | 9 |
| | | | | | | | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 119 2.4.1.24 | 10 |
| | | | | | | | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 7 4.2.2.13 | 4 4.2.2.13 | 4 4.2.2.13 | 11 |
| | | | | | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 17 2.4.1.24 | 12 |
| | | | | | 1790 3.2.1.199 | 1790 3.2.1.199 | 1790 3.2.1.199 | 1790 3.2.1.199 | 1790 3.2.1.199 | 1787 3.2.1.199 | 1782 3.2.1.199 | 1782 3.2.1.199 | 1782 3.2.1.199 | 1782 3.2.1.199 | 1782 3.2.1.199 | 1782 3.2.1.199 | 1782 3.2.1.199 | 13 |
| | | | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 67 3.2.1.22 | 14 |
| | | 616 3.2.1.20 3.2.1.84 | 616 3.2.1.20 3.2.1.84 | 616 3.2.1.20 3.2.1.84 | 616 3.2.1.20 3.2.1.84 | 616 3.2.1.20 3.2.1.84 | 616 3.2.1.20 3.2.1.84 | 615 3.2.1.20 3.2.1.84 | 575 3.2.1.20 3.2.1.84 | 575 3.2.1.20 3.2.1.84 | 575 3.2.1.20 3.2.1.84 | 574 3.2.1.20 3.2.1.84 | 569 3.2.1.20 3.2.1.84 | 542 3.2.1.20 3.2.1.84 | 542 3.2.1.20 3.2.1.84 | 542 3.2.1.20 3.2.1.84 | 542 3.2.1.20 3.2.1.84 | 15 |
| | | | | | | | | | | | | | | 27 | 27 | 27 | 27 | 16 |
| | | | | | | | | | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 37 | 36 | 17 |
| | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 528 3.2.1.217 | 18 |
| | 181 3.2.1.22 | 178 3.2.1.22 | 173 3.2.1.22 | 173 3.2.1.22 | 173 3.2.1.22 | 162 3.2.1.22 | 118 3.2.1.22 | 118 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 117 3.2.1.22 | 19 |
| | | | | | | | 44 3.2.1.22 | 43 3.2.1.22 | 43 3.2.1.22 | 43 3.2.1.22 | 43 3.2.1.22 | 43 3.2.1.22 | 33 3.2.1.22 | 31 3.2.1.22 | 31 3.2.1.22 | 20 3.2.1.22 | 20 3.2.1.22 | 20 |
| | 27 | 33 | 41 | 45 | 64 | 82 | 87 | 89 | 102 | 109 | 122 | 125 | 142 | 145 | 148 | 174 | 195 | UC |

The table includes the number of sequences for each subfamily along with functional information (EC numbers) for biochemically characterized members within each subfamily.

determined representative from each subfamily, if available. Of the 20 subfamilies, 15 contain at least one biochemically characterized member with defined substrate, and of these characterized subfamilies, 12 contain at least one member for which an experimentally determined 3D structure is available (Table 3). A common feature of all GH31 members is the $(\alpha/\beta)_8$-barrel fold of the catalytic domain (Fig. 3). The active site is located within the catalytic domain and exists as either a pocket or a cleft. The active site contains two aspartic acids (Asp/D) that act as nucleophile and general acid/base catalyst for hydrolases and transglycosidases, or nucleophile/general base and general acid for lyases. Each structural representative contains four domains: an N-terminal domain, a catalytic domain, and proximal and distal C-terminal domains. These four domains are the active module for all GH31 members (GH31 module). Certain GH31 proteins contain extra domains such as carbohydrate binding modules (CBMs). We provide hereafter an overview of principal features of each GH31 subfamily.

Upon selection of $10^{-125}$ as the most appropriate SSN *E*-value, the RaxML platform (22) was used to obtain a phylogenetic tree to assess the evolutionary relationships among GH31 subfamilies. A maximum likelihood phylogenetic tree was obtained with 100 bootstrap replicates, using 30 sequences from each subfamily, and for subfamilies with less than 30 proteins, all sequences were used. This phylogenetic tree

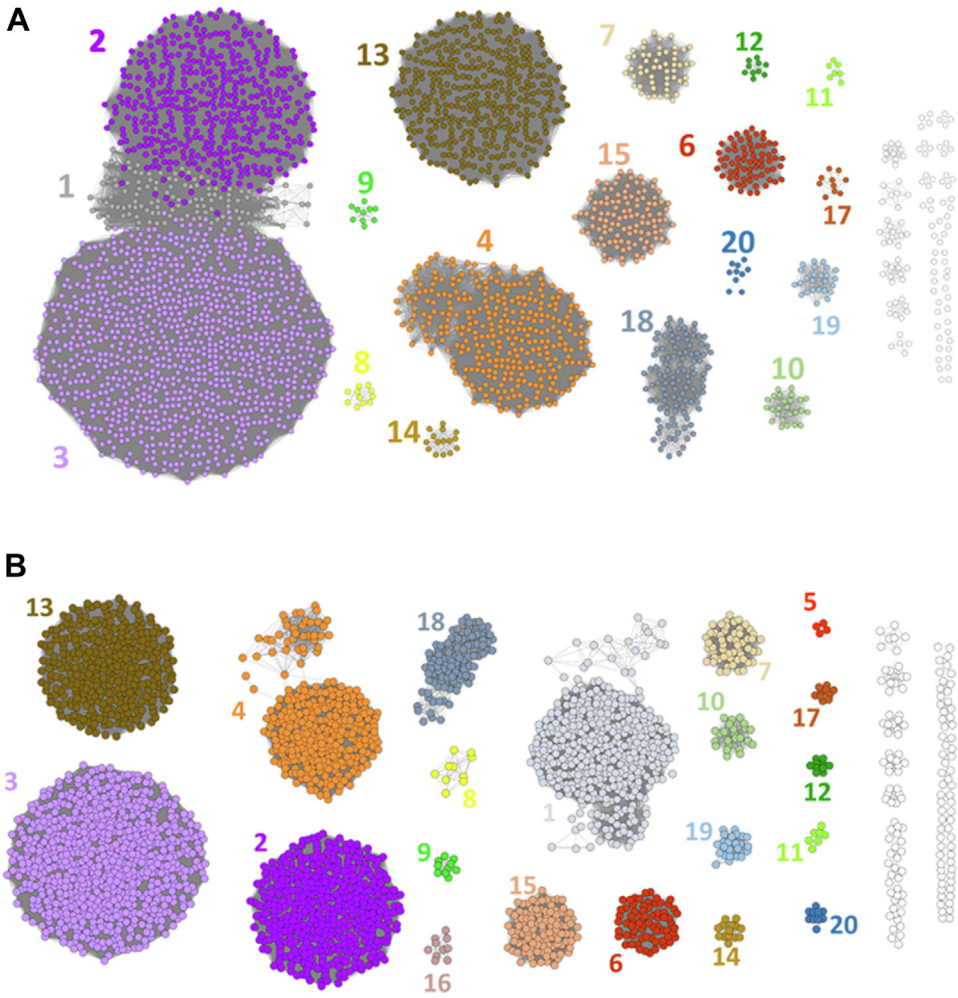


**Figure 2. Sequence Similarity Networks of GH31 sequences.** Numbers denote subfamily, GH31_*n*. *A*, for $E = 10^{-115}$ and *B*, for $E = 10^{-125}$ (networks were generated and annotated in Cytoscape). GH, glycoside hydrolase.

highlights evolutionary relationships between subfamilies, which are not readily apparent by SSN analysis because it lacks the protein evolution model of BLAST alignments.

### *Summary of characterized subfamilies*

*GH31_1*

One of the largest subfamilies, GH31_1 comprises 2092 enzymes from all kingdoms of life. It contains characterized members with nine different EC numbers and is thus the most functionally diverse subfamily, which could limit its predictive power. Ninety-four percent of characterized GH31_1 exhibit α-glucosidase activity, hydrolyzing α-(1→4), α-(1→6), or α-(1→3) linkages of various saccharides. The most widely distributed EC number in the subfamily is for α-glucosidase (3.2.1.20), with 84% for eukaryotic proteins and several from archaea and bacteria (mostly Firmicutes). The eukaryotic α-glucosidases include members from protozoa, fungi, metazoans, and plants. One archaeal protein exhibits both α-glucosidase and α-mannosidase (3.2.1.24) activities (14). Glucan 1,3-α-glucosidases (3.2.1.84) are found in subfamily 1 and are mostly from eukaryotes (23). Isomaltase (oligo-1,6-glucosidase, 3.2.1.10) activity in GH31_1 occurs in three members from metazoa and a Firmicutes bacterium. Subfamily GH31_1 includes three characterized plant α-xylosidases (3.2.1.177). These enzymes are localized to the apoplast and also possess α-glucosidase activity (16, 24). The subfamily also contains an oligosaccharide 4-α-glucosyltransferase (2.4.1.161, also termed amylase III), which catalyzes transfer of α-D-glucose residues from polysaccharides and maltooligosaccharides to create new α-1→4 linkages. Like many transglycosidases, the oligosaccharide 4-α-glucosyltransferase Agd31B from *Cellvibrio japonicus* in GH31_1 displays weak α-glucosidase activity, catalyzing the hydrolyses of the disaccharide maltose into glucose (25).

The structure of members of GH31_1 contains four major domains and two subdomains. The structural representative α-glucosidase from sugar beet (Protein Data Bank, PDB 3W37) has a long loop in the N-terminal domain that is a part of the active site pocket within the $(\alpha/\beta)_8$ barrel catalytic domain.

This subfamily contains several human glycosidases including α-glucosidase A (GAA, lysosomal), α-glucosidase II (GANAB), maltase-glucoamylase, and sucrase-isomaltase (SI). Maltase-glucoamylase and SI both duplicated GH31_1

## *Subfamily classification of GH family 31*

**Table 3**
**Subfamily classification of family GH31**

| EC code | Sub family No | Taxonomical diversity | Enzyme activity | PDB code |
|---|---|---|---|---|
| **2092**<br>3.2.1.10<br>3.2.1.11<br>3.2.1.20<br>3.2.1.24<br>3.2.1.48<br>3.2.1.84<br>3.2.1.177<br>2.4.1.24<br>2.4.1.161 | 1 | MUL | oligo-1,6-α-glucosidase<br>dextranase<br>α-glucosidase<br>α-mannosidase<br>sucrose α-glucosidase<br>glucan 1,3-α-glucosidase<br>α-D-xyloside xylohydrolase<br>1,4-α-glucan 6-α-glucosyltransferase<br>oligosaccharide 4-α-D-glucosyltransferase | 3W37<br>(11) |
| **2138** | 2 | MUL | α-glucosidase<br>sucrose α-glucosidase<br>glucan 1,3-α-glucosidase<br>α,α-trehalase | |
| **3573**<br>3.2.1.20<br>3.2.1.177 | 3 | MUL | α-glucosidase<br>α-D-xyloside xylohydrolase | 1XSI (2) |
| **1594**<br>3.2.1.177 | 4 | MUL | α-D-xyloside xylohydrolase | 2XVG<br>(4) |
| **4**<br>3.2.1.177 | 5 | Eukaryota | α-D-xyloside xylohydrolase | 6DRU<br>(1) |
| **333** | 6 | Proteobacteria | | |
| **267**<br>2.4.1.387<br>3.2.1.204 | 7 | Bacteria | isomaltosyltransferase<br>1,3-α-isomaltosidase | 5F7U<br>(3) |
| **24**<br>4.2.2.13 | 8 | MUL | exo-(1→4)-α-D-glucan lyase | |
| **26** | 9 | Bacteroidetes | | |
| **119**<br>2.4.1.24 | 10 | Bacteria | 1,4-α-glucan 6-α-glucosyltransferase | 5X7O<br>(1) |
| **7**<br>4.2.2.13 | 11 | MUL | exo-(1→4)-α-D-glucan lyase | 2X2H<br>(1) |
| **17**<br>2.4.1.24 | 12 | Bacteria | 1,4-α-glucan 6-α-glucosyltransferase | |
| **1782**<br>3.2.1.199 | 13 | MUL | sulfoquinovosidase | 5AED<br>(2) |
| **67**<br>3.2.1.22 | 14 | Bacteria | α-galactosidase | 4XPO<br>(1) |
| **542**<br>3.2.1.20<br>3.2.1.84 | 15 | MUL | α-glucosidase<br>glucan 1,3-α-glucosidase | 7WJ9<br>(1) |
| **27** | 16 | Bacteria | | |
| **40** | 17 | Bacteria | | |
| **528**<br>3.2.1.217 | 18 | MUL | exo-acting protein-α-N-acetylgalactosaminidase | 6M76<br>(1) |
| **117**<br>3.2.1.22 | 19 | Bacteria | α-galactosidase | |
| **31**<br>3.2.1.22 | 20 | Metazoa | α-galactosidase | 7QQF<br>(1) |

Taxonomical diversity: MUL, multiple kingdoms; PDB code: a selected structural representative for each subfamily for superposition and comparison among subfamilies, the number in brackets indicates the total number of characterized structures for each subfamily. Colors used in column 1 are the same as those for Table 2.

modules with all four modules individually exhibiting exo-glucosidase activities against linear α-1,4-linked maltose substrates but different preferences for oligosaccharides substrates with various lengths (26). The N-terminal GH31 domain of SI also has activity for the α-1,6-linkages of starch, and the C-terminal domain of SI has activity for the α-1,2-linkage of

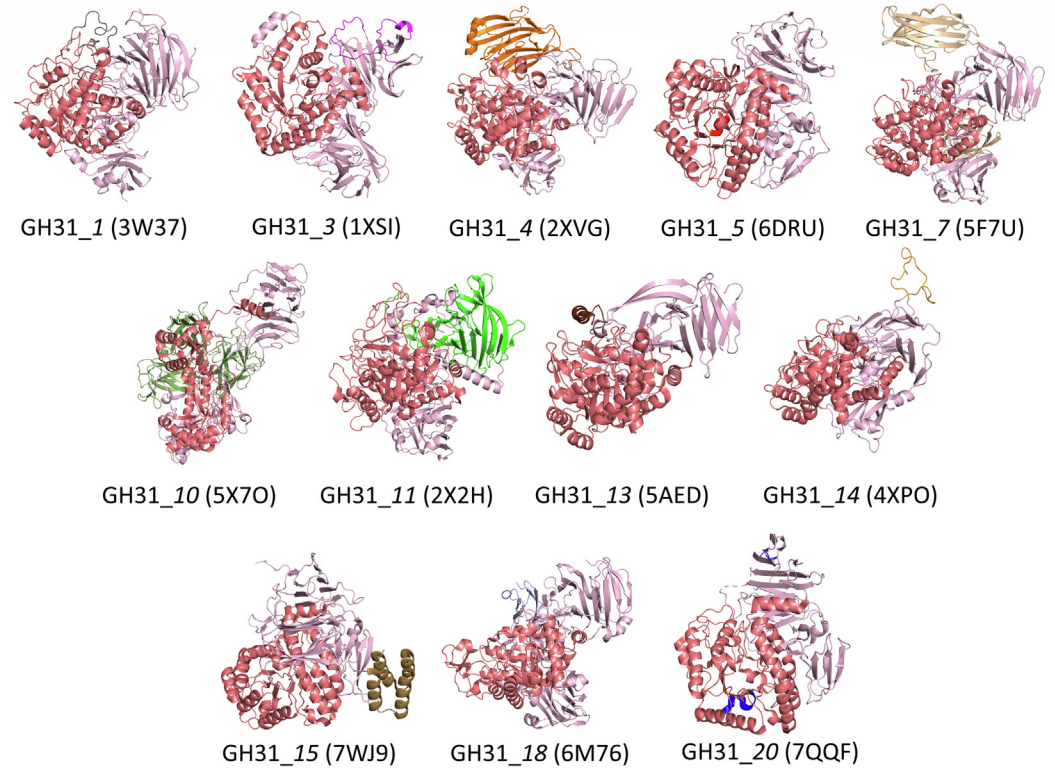


**Figure 3. 3D structures of subfamily representatives from the Protein Structure Databank highlighting common $(\alpha/\beta)_8$ catalytic domain in GH31 members.** The structures are oriented to highlight the common catalytic domain (shown in *deep salmon*) and the other conserved N- and C-terminal domains (in *light pink*). Other colors are used to highlight subfamily-specific domains, as discussed in the text. GH, glycoside hydrolase.

sucrose (27). The human hydrolases have an extra domain, an N-terminal trefoil type-p domain, which is a secondary substrate-binding domain (Fig. S2) (26, 28, 29).

### *GH31_3*

This large subfamily contains mainly bacterial sequences and some fungi and archaea. However, this may reflect sampling bias as the CAZy database contains mostly bacterial genomes (>15,000), with fewer archaeal (459), fungal (259), metazoan (29), and plant (23). Most bacterial members of this subfamily belong to gram-negative Gammaproteobacteria (>70% of Bacteria); the rest are mostly from the Terrabacteria taxon (Firmicutes and Actinobacteria). Just two activities are described for members of this subfamily, namely α-xylosidase (3.2.1.177) from five bacterial and fungal proteins and α-glucosidase (3.2.1.20) for a single Bacteriodetes protein.

There are subfamily structural representatives for an α-xylosidase and an α-glucosidase, with only slight differences between them. The α-glucosidase is a monomeric enzyme, and the active site is a cleft that binds small, linear maltooligosaccharides (30). The sole structure of an α-xylosidase (1XSI, YicI from *Escherichia coli*) from this subfamily is a hexamer, formed from a pair of trimers. This is the only known hexameric xylosidase in family GH31 (31, 32). A large loop in the N-domain, and the proximal C-domain, contribute to the hexameric structure. The active site is composed of residues from several of the monomers of the hexamer, suggesting that oligomerization of YicI may contribute to catalytic activity as well as stabilization of the protein fold. The two characterized α-xylosidases in GH31_3 have low α-glucosidase activity, and $Phe^{277}$ in YicI within the active site is located near C5 of the substrate and is suggested to be responsible for the low catalytic activity towards glucosides (31). YicI α-xylosidase shows outstanding activity on isoprimeverose (Xyl-α(1,6)-Glc), suggesting it is most likely active on xyloglucan. Superposition of α-glucosidases and α-xylosidases within this subfamily reveals that the latter has a longer distal C-domain (Fig. S3).

### *GH31_4*

This subfamily contains 1594 proteins from multiple kingdoms, with only one activity reported to date, α-xylosidase (3.2.1.177) for both bacterial and archaeal members. Most enzymes in the subfamily are from bacteria (96%) across various phyla (mostly Gammaproteobacteria, >50% of total bacterial sequences), with just 1.5% each from archaea and fungi. α-Xylosidases of GH31_4 are highly specific for xylogluco-oligosaccharides. Activity on α-glucosides are 6-fold lower and compared to α-xylosidases in GH31_3, activity on α-xylosides are 10-fold higher (33). There are four structurally characterized members in the family. A representative structure (PDB 2XVG, from *C. japonicus*) contains five domains (33). Four of the five are domains similar to other GH31 structures. However, there is an extra N-terminal domain that is specific to the subfamily that is classified as a PA14 domain (Fig. S1). PA14 is an all-β strand domain that facilitates

## Subfamily classification of GH family 31

binding of longer xyloside substrates within the active site cleft (33). Phylogenetic evaluation suggests that xylosidases of GH31_3 and GH31_4 are not closely related (Fig. 4), and structural superposition indicates that members in these two subfamilies have low structural similarity (RMSD > 2 Å), particularly because GH31_3 lacks the PA14 domain.

### *GH31_5*

This is a small subfamily that is a sister clade to subfamily 4 (Fig. 4). It contains a small number of ascomycotic fungal xylosidases (3.2.1.177). A structural representative (PDB 6DRU, from *Aspergillus niger*) has two asymmetric dimeric units that form a tetrameric biological unit (34) and lacks the PA14 of subfamily GH31_4 α-xylosidases. Tyr286 in the −1 substrate-binding site plays an important role in substrate specificity towards xyloglucan oligosaccharides (34). The bulky aromatic group in Tyr disfavors binding of C6 sugar substrates such as glucose and galactose in this site. A similar effect is seen in α-xylosidases of subfamily 4, whereas subfamily 3 members do not have a bulky aromatic residue at the corresponding position, leading to increased binding and hydrolysis of C6 carbohydrates.

### *GH31_7*

This is a bacterial subfamily with the majority (>70%) derived from pathogenic species. It contains two activities involved in synthesis or cleavage of cycloalternan (cyclobis-(1,6)-α-nigerosyl): cycloalternan-forming enzymes (3-α-isomaltosyltransferase/CAFE, 2.4.1.387) and cycloalternan-degrading enzymes (α-isomaltosidases/CADE, 3.2.1.204). Most members (98%) belong to two phyla of grampositive bacteria in the Terrabacteria taxon, namely Firmicutes (73% of total bacterial population) and Actinobacteria.

There is a single 3D structure of a CAFE and 3D structures of two CADEs. All are broadly similar, with the similarity extending to the active sites, which may reflect the fact that both CAFEs and CADEs bind cycloalternan (as product and substrate, respectively). However, the structures exhibit several key differences. The structure of CAFE (PDB 5F7U, from *Listeria monocytogenes*) consists of six domains, including the GH31 catalytic module, an N-terminal CBM-like domain, and a C-terminal CBM domain. The C-terminal CBM domain belongs to CBM35 family based on sequence homology, and it is proposed to enhance transglucosylation activity (35, 36). This trimodular CBM-GH-CBM architecture is frequent in the subfamily (∼30%), across Terrabacteria. The representative CADE (PDB 5F7S, from *Trueperella pyogenes*) has four domains but lacks the C-terminal CBM and distal C-terminal domain (CD2). The RMSD value between these two structures are close to 1 Å (Fig. S4). A β3-β4 insertion between the N-terminal and CBM-like domain of CADE is disordered in the ligand-free state but becomes ordered upon binding of cycloalternan (37). This β insertion is not present in the CAFE sequence. In addition, the α4-α5 loops of CAFE and CADE proteins have distinct conformational behaviors (Fig. S4). In ligand-free CADE, the α4-α5 loop has a conformation that creates a shallow active site that transforms to a deep conformation upon binding of cycloalternan. By contrast, the same loop in CAFE exhibits the deep conformation in both unliganded and cycloalternan-bound states. This appears to favor the transglucosylation activity of CAFE (36, 37).

### *GH31_8*

This subfamily, along with GH31_11 is one of two that contains exclusively α-glucan lyases (4.2.2.13). Phylogenetic analysis shows GH31_8 is a sibling to GH31_11 (Fig. 4).

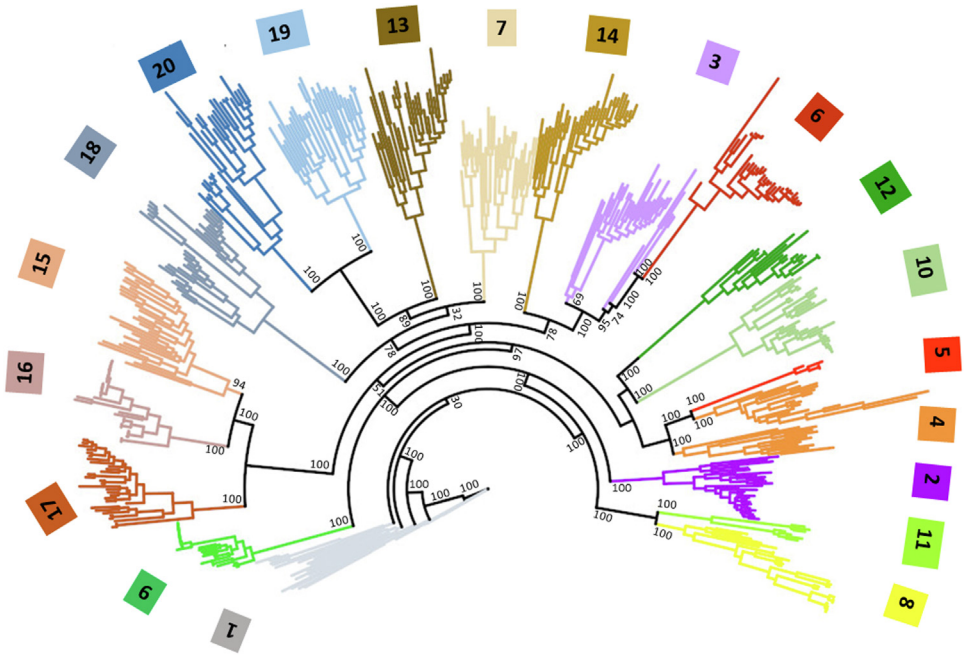


**Figure 4. Phylogenetic tree for GH31 subfamily distribution.** Bootstrap values are indicated on the node of each branch. GH, glycoside hydrolase.

Members are mainly from fungi (Ascomycota and Basidiomycota) and include a few from Cyanobacteria and Deltaproteobacteria. The number of sequences remain the same across a range of *E* values (from $10^{-110}$ to $10^{-150}$) used for SSN construction, emphasizing the cohesion of the group. This subfamily does not have any structural representative.

### GH31_10

This small subfamily contains 119 members, with four characterized members with a sole activity, 6-α-glucosyltransferase (6GT) (2.4.1.24). They catalyze the transglycosylation of α-(1→4)-glucosidic linkages to form a nonreducing terminal (1→6)-α-D-glucose linkage and liberate glucose. All members are from bacteria, with most from Firmicutes (Clostridia), Actinobacteria, and several Proteobacteria clades. The four characterized members in the family are Firmicutes and exhibit transglucosylation and apparent hydrolysis activities.

The subfamily contains one representative with a 3D structure (PDB 5X7O, from *Paenibacillus* sp. 598K) (38). It has three extra C-terminal β-jellyroll domains compared to other GH31 members (Fig. S1). These three β-jellyroll domains are CBMs, two of the domains belong to the CBM35 family, the other to the CBM61 family. This architecture is prevalent in the Bacilli, unlike Gammaproteobacterial members which consist of only the GH31_10 module. Based on structural data for 3D structures in complex with isomaltose or isomaltotriose, the binding preference of the first CBM35 was assigned as α-1,6-glucan, the second CBM35 was assigned as binding α-1,4-glucan based on a maltooligosaccharide complex, and the CBM61 was assigned as binding α-1,4-glucan based on complexes with α-1,6-glucosyl-maltotriose and acarbose (38). The $(\alpha/\beta)_8$-barrel catalytic domain is compactly surrounded by the three CBMs, which are proposed to accelerate the catalytic efficiency and avidity of binding. The N-terminal domain exhibited a binding preference for α-1,4-glucan (38). Loops 6 and 7 in the catalytic domain are not well conserved with the other GH31 subfamilies. The GH31_10 6GTs have a wider catalytic cleft in comparison to other GH31 hydrolases, which may promote transglucosylation. Structural superposition indicates that the GH31_10 member has low structural similarity to most structurally characterized GH31 members (primarily RMSD > 5–20 Å).

### GH31_11

The GH31_11 subfamily contains seven members, with four characterized members that are all α-glucan lyases (4.2.2.13). The characterized members are all from red algae, and the subfamily contains other members from Alphaproteobacteria and Gammaproteobacteria.

Only one 3D structure (PDB 2X2H), an α-glucan lyase from the red seaweed *Gracilariopsis lemaneiformis*, is available for this family. This α-glucan lyase has the same number of domains as GH31 α-glucosidases, and a starch-binding site was identified in the N-terminal domain, which acts as a secondary carbohydrate-binding site. The structure has two aspartic acids ($Asp^{553}$ and $Asp^{665}$) that act as the nucleophile and acid catalyst, resembling other GH31 members. However, the unique feature that governs the α-glucan lyase activity is the nucleophile ($Asp^{553}$), which acts as an intramolecular base in the second step of the reaction mechanism to abstract the C2 proton within the covalent β-glucosyl enzyme bound in the −1 subsite in the elimination reaction. Val at position 556, instead of Glu seen in the other GH31 subfamilies, promotes lyase activity (18). The *Lonsdalea britannica* bacterial α-glucan lyase has Ala instead of Val at the equivalent position (Fig. S5) and may play a similar role.

As noted above, phylogenetic analysis (Fig. 1.3) suggests that subfamilies GH31_8 and GH31_11 are closely related sibling subfamilies that may have evolved from a common ancestor, even though they differ in their taxonomic distribution and separate very early in the SSN analysis, suggesting low sequence similarity levels. Thus, members of GH31_8 may have similar structures to α-glucan lyases in subfamily GH31_11. Position 556 in the *G. lemaneiformis* enzyme, which is believed to influence lyase activity, is Val/Ala in GH31_11. On the other hand, members of subfamily 8 exhibit greater diversity at the equivalent position, with residues including Thr (in most), Asp, Cys (only in one), Gly, or Ser. Superposition of the AlphaFold (39, 40) predicted structure for subfamily 8 (Fig. S8) and experimentally determined subfamily 11 structures yields RMSD of 1.2 Å.

### GH31_12

This is a small subfamily of just 17 sequences that contains 6GT activity. A small amount of hydrolytic activity was seen on maltooligosaccharides (41). The members are all of bacterial origin (mostly from Actinobacteria and Gammaproteobacteria along with two from Cyanobacteria). Its closest phylogenetic connection is with GH31_10 (Fig. 4). Like subfamily 10, most subfamily 12 members are multimodular, commonly with a CBM20 domain (which are associated with starch and cyclodextrin binding).

Both subfamilies 10 and 12 are involved in the transglucosylation/rearrangement of α-glucan to give α-D-isomaltosyl-(1→4)-α-D-glucan, the first step in the synthesis of cycloalternan (41, 42). These transglucosidases have catalytic residues (Asp) that are conserved with GH31 α-glucosidases (41).

The AlphaFold GH31_12 predicted structure, BAD34980.1 from *Arthrobacter globiformis*, has a β-jellyroll CBM domain (Fig. S8) and has low structural similarity with the other subfamily of 6GT (RMSD ∼28 Å).

### GH31_13

This subfamily contains sulfoquinovosidases (3.2.1.199), which are exo-acting enzyme that catalyze cleavage of glycosides of the sulfosugar 6-deoxy-6-sulfo-D-glucose (43). This subfamily emerges early in the SSN analysis. A wide range of taxonomic diversity is seen within the subfamily, including members of bacteria (98%), fungi, algae, plants, and metazoans. Bacterial members belong to various phyla, with the majority belonging to the Gammaproteobacteria (94% of total bacterial

## Subfamily classification of GH family 31

population) and Alphaproteobacteria. Consistent with the importance of sulfoquinovose in the function of the thylakoid membrane, most eukaryotic members are from photosynthetic organisms. Several members are from metazoa including the marine tunicates *Phallusia mammillata* and *Oikopleura dioica* and crustaceans *Cyprideis torosa* and *Darwinula stevensoni*.

Three 3D structures of SQases (5AED, 5OHY, and 6PNR) have been reported and are highly similar, consisting of an $(\alpha/\beta)_8$-barrel catalytic module but with the distal C-domain shorter than most family GH31 hydrolases. The sulfonate group of sulfoquinovose in the −1 subsite makes interactions with three residues (Arg/Trp/Tyr; RWY) of the enzyme directly or indirectly (*via* a bridging water molecule hydrogen-bonded to Tyr). Sequence alignment of GH31_13 indicates that most bacteria possess QQRWY and KERWY motifs, which has been studied in detail (17, 44). QQWY/QQWF motifs are present in other bacteria with the former in some plants. A fungal representative possesses PRWY, and in a previous study (17) metazoan members of this subfamily were reported to possess QRWF/QRWY motifs (Fig. S6).

### *GH31_14*

This subfamily contains bacterial α-galactosidases (3.2.1.22). The majority are from Bacteriodetes (82%), while the rest includes members of Firmicutes (12%) and a member each from diverse phyla Acidobacteria, Verrucomicrobia, Gemmatimonadetes, and Lentisphaeria.

There is a single subfamily structural representative (4XPO, from *Pedobacter saltans*) that contains four main domains, like other GH31 members (45). However, the structure reveals an inserted loop in the N-domain. This loop is involved in the formation of a dimeric biological unit and influences α-galactosidase activity as it facilitates the aglycon specificity and enables binding of L-fucose. Another unique loop is inserted in the C-domain, but in the 3D structure, it is disordered. The −1 subsite residues are unique from other characterized structures in GH31 family. The enzyme binds L-fucose in the +1 subsite. However, its substrate specificity differs to other characterized α-galactosidases as no hydrolysis activity was detected on various natural galactosides such as galactomannans, suggesting that subfamily members may be involved in a novel α-D-galactopyranosyl-L-fucose degradation system (45).

### *GH31_15*

This subfamily contains 542 sequences from a wide taxonomic diversity that includes archaea, bacteria, and eukaryotes. Bacterial members are mostly from the Terrabacteria taxon (Firmicutes (>52%) and Actinobacteria (>22%)) and Bacteroidetes. The eukaryotic members are mainly from Fungi (>99%) but includes a few members from red algae and Oomycota. Members of the subfamily catalyze hydrolysis of the α-1,3-linkage of glucosides (3.2.1.84) and α-1,4-linkages of glucosides (3.2.1.20). Experimental data shows a preference for hydrolysis of α-1,3 linkage of nigerooligosaccharides including nigerose (46).

The structural representative (PDB 7WJ9, from *Lactococcus lactis* subsp. *cremoris*) of the subfamily is a hexamer with each protomer consisting of five domains (46). Four are the conserved GH31 domains, while the fifth, a C-terminal α-helix domain containing four helices, is involved in formation of the hexamer. This enzyme has very low activity on maltooligosaccharides, which is ascribed to a lack of space in the +1 subsite of the catalytic pocket (46).

### *GH31_18*

This is the only family GH31 subfamily with members active on a 2-acetamido sugar, namely exo-acting protein α-*N*-acetylgalactosaminidase (3.2.1.217) (19). The subfamily spans several kingdoms but consists mainly of bacterial sequences from several phyla (mostly Firmicutes, Bacteroidetes, Actinobacteria, and Gammaproteobacteria) and just a few metazoan members. The cohesion of this subfamily is evidenced by its early emergence at low stringency ($E < 10^{-60}$) and it does not disaggregate even at $E = 10^{-140}$.

In the 3D structure (PDB: 6M76) from *Enterococcus faecalis*, an extra subdomain is present in the catalytic domain that includes a short α-helix and an antiparallel β-sheet (47). There is a conformational change of the active site between the ligand-free (open) and ligand-bound (closed) forms (47), which has not been observed in other GH31 subfamily members, and may be connected to the unique exo-acting *N*-acetylgalactosaminidase activity of this subfamily. There is an extra domain in the polypeptide, a fibronectin type 3 domain. Fibronectin type 3 domains can function as a linker connecting catalytic and CBM domains (48). The reported structure does not include a CBM domain, but the gene was truncated for expression, and the full length protein has a CBM32 domain like the majority of members of this subfamily (others contain CBM32 and CBM51 domains) (47). The hydrolytic activity of GH31 *N*-acetylgalactosaminidases is independent of the C-terminal domain.

### *GH31_19*

This subfamily contains 117 bacterial proteins including members with α-galactosidase activity (EC 3.2.1.22) (49). The majority of the group contains members from bacterial phyla, mainly Bacteroidetes (47%), Firmicutes (38%), and Actinobacteria (15%).

A sequence alignment of subfamily 14 (another α-galactosidase subfamily) and 19 was conducted (Fig. S7). The alignment indicates differences around the N-domain region, especially in the inserted loop and catalytic domain, which may result in structural differences.

Structural superposition of the AlphaFold predicted GH31_19 structure (Fig. S8) and the GH31_14 experimentally determined structure gives RMSD of 2.3 Å, with significant differences in N- and C-terminal domains.

### *GH31_20*

This small subfamily contains only metazoan members (31) with one characterized as an α-galactosidase, myogenesis-

regulating glycosidase, MYORG, from *Homo sapiens* (50). The majority of members belong to phylum Arthropoda (>60%), and the rest are from Chordata and Annelida. The GH31_20 subfamily is a sibling to subfamily 19 and formed a single group at $E = 10^{-90}$.

The subfamily contains a single 3D structure (PDB 7QQF) of MYORG (50). MYORG forms a dimer in solution and crystals. It has three main GH31 domains, except lacks the usual distal C-terminal domain, and has a disordered region in N-terminal domain. It has an insertion (between α3 and α4) in $(\alpha/\beta)_8$-barrel catalytic domain. It is proposed that Trp321 in the −1 subsite is responsible for α-galactosidase rather than α-glucosidase activity. Despite being a human enzyme, MYORG lacks activity on α-galactose containing disaccharide structures that exist in humans including Gal-α1,3-Gal, Gal-α1,3-GalNAc, and Gal-α1,4-Gal but has excellent activity on the unusual Gal-α1,4-Glc disaccharide and weak activity on Gal-α1,6-Gal. Asp213 and Arg504 in subsite +1 are engaged in hydrogen bonding interactions with the glucose of Gal-α1,4-Glc. Trp426 in the +1 subsite participates in stacking interactions with glucose.

Subfamily 14 is an α-galactosidase subfamily with a structural representative. Structural superposition of members of subfamily 14 and 20 has RMSD of 2.5 Å. Both share several active site residue similarities to encourage α-galactosidase activity, however, the subfamily 14 structural representative has Glu366 instead of Trp426, indicating a different substrate preference.

Sibling subfamily 19, another α-galactosidase subfamily, lacks a structurally characterized representative; however, superposition of the AlphaFold predicted structure of subfamily 19 (Fig. S8) and MYORG structure are similar with RMSD of 1.5 Å.

### *Unclassified sequences*

The SSN classification scheme leaves only 1% of GH31 modules unclassified into subfamilies (for a detailed description of the minimum threshold for defining a family, see Experimental procedures). Most of the unclassified modules are singletons that do not share similarities to any other family member. There are several small clusters of proteins that lack the minimum number of proteins to define a subfamily and do not contain characterized members. It is expected that these groups may integrate subfamilies or become new subfamilies as more sequences are deposited and curated into the CAZy database.

### *Uncharacterized subfamilies*

Our GH31 subfamily classification has six subfamilies without any characterized member and highlights the sequence space of the GH31 family remaining to be explored to improve its predictive power. GH31_2 is one of the largest subfamilies and exhibits high taxonomic diversity. It contains 2138 members including PGH31 from *Paecilomyces lilacinus*, which is implicated in the nematophagous action of this fungus in biocontrol of oilseed rape (51) but no biochemical activity has been reported for any member yet. We therefore chose to characterize one member of this subfamily (vide infra; see next section).

Subfamilies 16 and 17, which lack characterized members, are closely related to subfamily 15 and they form a single group at $E = 10^{-65}$ (Table 2 and Fig. 4). GH31_17 splits from the group at $E = 10^{-100}$ and contains 40 bacterial sequences (95% Actinobacteria). At $E = 10^{-125}$, the remaining group splits into two, forming subfamilies GH31_16 and GH31_15. Subfamily GH_16 contains only 27 members that are exclusively gram-negative bacteria (96% Bacteroidetes). This splitting therefore appears to be a taxonomical separation from subfamily 15. Whether subfamilies 16 and 17 also display specificity for nigero-oligosaccharides is an open question that awaits experimental study. The topology of the phylogenetic tree indicates that these three subfamilies are sibling clades in a separate triple rooted branch (Fig. 4) and suggests that subfamilies 16 and 17 may have similar activities to subfamily 15 (EC 3.2.1.84). Structural superposition of AlphaFold predicted structures of representatives of subfamilies 16 and 17 (Fig. S8) with experimentally determined subfamily 17 gives RMSD 1.3 Å, but subfamily 17 has a longer chain. Furthermore, subfamilies 16 and 17 lack the C-terminal α-helix domain in subfamily 15, and subfamily 17 may have an extra domain in the C-terminal domain.

The Proteobacteria GH31_6 and Bacteroidetes GH31_9 subfamilies also lack characterized members. Subfamily 6 is formed at $E = 10^{-115}$ along with subfamily 4, which consists of α-xylosidase members from subfamily 1. Phylogenetic analysis suggests that subfamilies 6 and 3 (which also includes α-xylosidase) evolved from a common ancestor (Fig. 4). The structural superposition of the AlphaFold predicted GH31_6 and GH31_3 structures (Fig. S8) has RMSD of 1.9 Å, and most parts of the catalytic, C-terminal and N-terminal domains overlay with each other. Collectively, this suggests that subfamily GH31_6 may also have α-xylosidase activity. However, the GH31_6 AlphaFold predicted structure lacks the large loop in the N-domain present in GH31_3. Subfamily 9 first appeared at $E = 10^{-105}$ and split from subfamily 1. However, no subfamily exhibits a close evolutionary relationship with GH31_9. The AlphaFold predicted structure of GH31_9 subfamily member QJR56767.1 from *Phocaeicola doreihas* displays a four domain GH31 module (Fig. S8).

### *Subfamily GH31_2 contains a broad-spectrum α-glucosidase*

The gene encoding the GH31_2 subfamily member *Ra*GH31 (QDL53937.1) from *R. aquaticus* sp. nov. isolate Gr-4$^T$ (GenBank: CP036282.1) was synthesized in a codon optimized form for *E. coli*, expressed in this host and purified to homogeneity. The enzyme was initially screened against a range of chromogenic substrates. The enzyme was active against 4-nitrophenyl α-D-glucopyranoside (PNPGlc) but was inactive against other α-glycosides including 4-nitrophenyl α-D-sulfoquinovoside (PNPSQ), 4-nitrophenyl α-D-galactopyranoside (PNPGal), 4-nitrophenyl α-D-xylopyranoside (PNPXyl), 4-nitrophenyl α-D-mannopyranoside (PNPMan), 4-nitrophenyl *N*-acetyl-α-D-galactosaminide (PNPGalNAc), 4-

nitrophenyl α-D-glucuronide (PNPGlcA), and 4-nitrophenyl α-D-glucopyranoside 6-phosphate (6-P-PNPGlc), representing the head-group of the majority of activities in family GH31. $^1$H NMR of the initially formed product revealed the formation of α-glucose, and thus that the enzyme is a retaining α-glucoside hydrolase (and not an α-glucan lyase). The effects of pH on the hydrolytic activity and temperature stability test of *Ra*GH31 were evaluated using PNPGlc as a substrate (Fig. 5). The pH dependence on $k_{cat}/K_M$ gave a bell-shaped curve, with an optimum at pH 6 and p$K_{a1}$ and p$K_{a2}$ values of 5.9 ± 0.4 and 5.7 ± 0.4 (Fig. 5*A*). The closeness of these estimated macroscopic p$K_a$ values may represent an example of 'reverse protonation', in which the microscopic p*K*a values are reversed in order with respect to the (apparent) macroscopic p*K*a values, and thus only a small fraction of enzyme is in the correct ionization state for catalysis (52). Kinetic parameters for PNPGlc measured at pH 6 were $k_{cat}$ = 1.55 $s^{-1}$, $K_M$ = 0.18 mM, and $k_{cat}/K_M$ = 8670 $M^{-1}$ $s^{-1}$ (Fig. 5*B*)

Hydrolytic activities toward various disaccharides was qualitatively assessed by TLC: trehalose [α-D-Glc*p*-(1↔1)-α-D-Glc*p*], kojibiose [α-D-Glc*p*-(1→2)-D-Glc], nigerose [α-D-Glc*p*-(1→3)-D-Glc], maltose [α-D-Glc*p*-(1→4)-D-Glc], isomaltose [α-D-Glc*p*-(1→6)-D-Glc], and sucrose [β-D-Fru*f*-(2↔1)-α-D-Glc*p*]. After 5 h incubation with 12.3 μM *Ra*GH31, the reactions were analyzed by TLC. Sucrose was completely hydrolyzed; partial hydrolysis of maltose, nigerose, and trehalose was also observed; isomaltose and kojibiose were inert (Fig. 5*C*). The kinetic parameters for hydrolysis of sucrose, maltose, nigerose, and trehalose were assessed using a colorimetric glucose detection assay. Kinetic parameters ($k_{cat}/K_M$) reveal a mild preference for sucrose over maltose and nigerose, while trehalose was the poorest of the four disaccharide substrates (Fig. 5, *D*–*G* and Table 4). The AlphaFold 3D structure prediction of *Ra*GH31 shows excellent structural similarity to an experimentally determined X-ray structure of *Chaetomium thermophilum* GIIα-D556A mutant with nigerose (PDB: 5DKZ, green, a member of subfamily 1) (Fig. S9).

## Discussion

SSNs provide a computationally efficient approach for analysis of sequence relationships in large protein families (53) and has previously been applied for classification of other GH families (10, 11). A key advantage of all-*versus*-all pairwise BLAST for SSN analysis is that it avoids the computationally intensive multiple sequence alignment which remains computationally challenging even after down-sampling, and for which the huge sequence diversity may lead to difficulties in identifying phylogenetic signal needed for assembling a robust phylogeny. We applied SSNs to divide 13,473 GH31 sequence modules into 20 subfamilies.

In this work, we followed the framework established by Viborg et al. who used BLAST-SSN to develop the subfamily classification of family GH16 (10). Analysis of precision and recall, taxonomic considerations, separation of distinct substrate preferences, analysis of catalytic mechanism, and the search for "reasonable-size" subfamilies guided our selection. In the case of GH31, the distribution of EC numbers was useful for defining a meaningful *E*-value threshold, with careful monitoring to ensure that excessive taxonomic fractionation did not occur or could be justified by structural or observable constraint in the subfamily alignments.

Family GH31 is unusual among the various GH families as it possesses members that use three different mechanisms—the classical retaining glycosidases, transglycosidases, and the α-glucan lyases. In addition, the family contains enzymes that act on diverse substrates: variation of substituents at the C5 position (H, xylosides; $CH_2OH$, glucosides; $CH_2SO_3^-$, sulfoquinovosides; $CO_2H$, glucuronosides), the C4 position (D-*gluco* and D-*galacto* configuration), and the C2 position (OH, D-*gluco*/D-*manno*; NHAc, *N*-acetylgalactosaminides). Generally, there was excellent partitioning of these variations into separate subfamilies, which might reflect the amino acid residue changes required in the active site to achieve these distinct mechanistic and substrate binding variations, as well as evolutionary divergence from common ancestors.

At the preferred *E*-value of $10^{-125}$ for subfamily classification, GH31 divides into 20 subfamilies, with 11 possessing just one EC number, three with two EC numbers, and one (GH31_1) possessing 9. In the case of subfamily GH31_1, some of these activities are overlapping and may even be shared by the same enzyme: dextranase and oligo-1,6-α-glucosidase; sucrose α-glucosidase and α-glucosidase; and α-mannosidase and α-glucosidase. Moreover, there is a close relationship between certain transglycosidases and glycosidases when they act on the same substrate and lead to either transfer to water (glycosidase) or transfer to a sugar acceptor (transglycosidase) *via* a common glycosyl enzyme intermediate. Examples include isomaltosyltransferase and 1,3-α-isomaltosidase, 1,4-α-glucan 6-α-glucosyltransferase oligosaccharide 4-α-D-glucosyltransferase, and dextranase.

Just five subfamilies had no EC number assigned. Subfamily GH31_2 is the largest of these uncharacterized subfamilies. Using a range of PNP glycosides corresponding to most monosaccharide variations observed within this family, we showed a GH31_2 family member from *R. aquaticus* possesses α-glucosidase activity (and not α-glucan lyase activity), as well as activity on several disaccharides including sucrose, nigerose, maltose, and trehalose, defining a series of EC numbers for this family.

The subfamily classification allows a systematic survey of the protein fold for structurally characterized members of family GH31. All experimentally determined members contained a conserved GH31 $(\alpha/\beta)_8$ barrel catalytic domain and three additional domains: an N-terminal domain, and proximal and distal C-terminal domains. However, in addition to these conserved domains, there is considerable structural diversity: human hydrolases in subfamily GH31_1 have an N-terminal trefoil type-p domain; subfamily GH31_4 contains an N-terminal PA14 domain; CAFE from subfamily GH31_7 contains an N-terminal CBM-like domain and a C-terminal CBM35 domain; subfamily GH31_10 contains three extra C-terminal β-jellyroll domains; subfamily GH31_15 contains a C-terminal α-

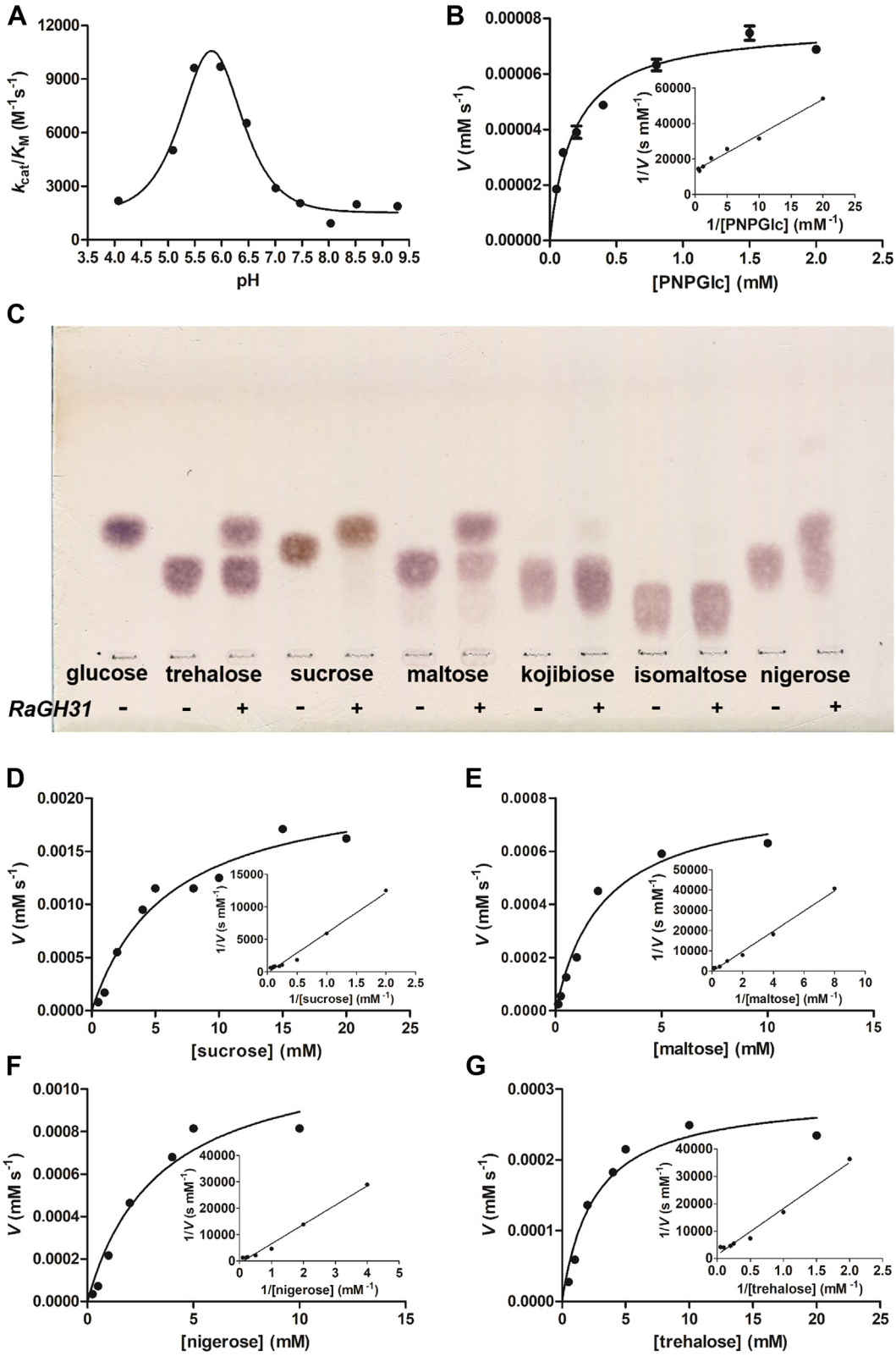


**Figure 5. Subfamily GH_2 *Ra*GH31 from *Rhodoferax aquaticus* sp. nov. isolate Gr-4$^T$ is an α-glucosidase with activity on trehalose, sucrose, maltose, and nigerose.** *A*, pH dependence of $k_{cat}/K_M$ using PNPGlc as substrate, monitored using a UV/vis spectrophotometer. *B*, Michaelis–Menten plot for PNPGlc as substrate. (inset) Lineweaver-Burk plot. *C*, HPTLC analysis of *Ra*GH31 digests of various disaccharides. Products were visualized with 0.2% orcinol in 10% $H_2SO_4$, 10% $H_2O$ in ethanol. Michaelis–Menten and Lineweaver–Burk (inset) plots for (*D*) sucrose, (*E*) maltose, (*F*) nigerose, and (*G*) trehalose, monitored using a coupled assay. GH, glycoside hydrolase.

**Table 4**
**Michaelis–Menten kinetic parameters for subfamily GH_2 *Ra*GH31 from *Rhodoferax aquaticus* sp. nov. isolate Gr-4$^T$**

| Substrate | $K_M$ (mM) | $k_{cat}$ ($s^{-1}$) | $k_{cat}/K_M$ ($M^{-1}$ $s^{-1}$) |
|---|---|---|---|
| PNPGlc | 0.17 ± 0.02 | 1.56 ± 0.04 | 8700 ± 960 |
| sucrose | 4.0 ± 1.1 | 0.18 ± 0.02 | 45 ± 13 |
| maltose | 2.2 ± 0.6 | 0.062 ± 0.006 | 28 ± 8 |
| nigerose | 3.4 ± 1.2 | 0.091 ± 0.014 | 27 ± 11 |
| trehalose | 2.6 ± 0.7 | 0.022 ± 0.002 | 8.6 ± 2.5 |

helix domain of four helices; and subfamily GH31_18 contains short α-helix and an antiparallel β-sheet domain.

## Conclusion

We describe a subfamily classification of the mechanistically and functionally diverse family GH31 using SSN analysis. This classification collates sequence, biochemical, mechanistic, and structural data on characterized members and supports more refined bioinformatic predictions, AlphaFold predictions, and provides a guide for experimental studies. The identification of subfamilies and unclassified sequences with no functionally characterized members offers an opportunity for future enzyme discovery. To support future sequence annotation and experimental design, the GH31 subfamily classification is now publicly available in the CAZy database (http://www.cazy.org/GH31.html).

## Experimental procedures

### Data acquisition

13,464 GH31-containing Genbank sequences were extracted using the SACCHARIS Perl script *cazy_extract.pl* (https://github.com/DallasThomas/SACCHARIS) based on CAZy database lists (June 2021) (54). The GH31 module was manually defined for biochemically characterized GH31 members on the basis of multiple sequence alignment in the program MAFFT using the G-INS-I strategy. This was used to conduct an *hmmsearch* in HMMER3-3 to identify the GH31 modules for all family members. Sequences shorter than 40% of query/reference modules were eliminated from the SSN analysis (0.03 in total%) resulting in 13,473 GH31 modules, saved as a FASTA-format file.

### SSN analysis

13,473 GH31 modules were introduced to the SSNpipe tool (10), with default settings to obtain all-*versus*-all pairwise local alignments of all GH31 modules by BLAST+ 2.2.31 (55). BLAST *E*-values ranging from $10^{-60}$ to $10^{-140}$ by steps of $10^{-5}$ allowed the definition of a series of 17 SSNs. In each SSN, every connected component (set of nodes connected to each other by any path) was considered as a candidate subfamily if it either (i) contained at least one characterized member and sufficient sequence diversity (15 proteins for prokaryotes, 4 for eukaryotes - less populated in CAZy due to incompleteness of eukaryotic assemblies) or (ii) contained at least 20 proteins. The remaining sequences were considered unclassified (uc). A Python script was used to down sample the number of nodes for an SSN to reduce the number of edges. SSNs were visualized using *yFiles organic* layout in Cytoscape-3.90.

### Performance analysis using HMM

For each SSN *E*-value threshold, a library of HMMs for the corresponding GH31 subfamilies, along with one extra HMM for the unclassified GH31 sequences were generated using the four following steps. Each sequence set (subfamily or the unclassified group) was subjected to the sequence redundancy reduction using the CD-hit (56) online platform with default parameters except a clustering percentage tuned to 75%. The resulting low-redundancy sets were aligned using the G-INS-i strategy in MAFFT (57). HMMs were built using the *hmmbuild* command in HMMER-3.3 (58) software after defining boundaries for each multiple sequence alignment using Jalview. The *cat* bash command in HMMER-3.3 was used to concatenate all HMM profiles (all subfamilies plus the uncharacterized) into a single HMM library for each SSN *E*-value.

The HMM library was used to predict subfamilies in the 13,473 GH31 modules. Hits were obtained using the *hmmscan* command from HMMER3-3 software with default parameters. The modules were assigned to a subfamily based on two aspects: a minimal *per-domain* HMM *E*-value of $10^{-100}$ and a required margin between the first/assigned and second-best hits of $10^{-20}$. Using a confusion matrix, true positives, false positives, and false negatives were calculated for each SSN *E*-value and used to deduce the overall precision and recall for each SSN *E*-value.

### Phylogenetic analysis

Thirty sequences for each subfamily were randomly selected. All sequences were taken from the subfamilies with less than 30 members. Each sequence set was then aligned with MAFFT using the G-INS-i strategy. A maximum likelihood phylogenetic tree was obtained from RAxML (22) with 100 bootstrap replicates. iTOL: Interactive Tree of Life (59) was used to visualize and annotate the best tree.

### Structural comparison

Twenty-nine PDB structures were taken from the PDB. One structural representative was chosen for each subfamily (if available). Multiple superposition for each subfamily was conducted using the *super* command in PyMOL-2.3.4. For subfamilies that exhibit two or more activities, the structures for each activity (if available) were compared using the same command in PyMOL-2.3.4.

### Cloning, expression, and purification of RaGH31

A dsDNA oligonucleotide encoding QDL53937.1 (*Ra*GH31) and codon-harmonized for expression in *E. coli* was synthesized (IDT Genscript) and cloned into the pET29b(+) (Novagen) expression vector using the *Nde*I/*Xho*I restriction sites to give the pET29-*Ra*GH31 plasmid (Fig. S10).

For protein expression, pET29-*Ra*GH31 was transformed into chemically competent 'T7 Express' *E. coli* cells (NEB) and

transformants selected on LB-agar (50 μg $ml^{-1}$ kanamycin) by incubation at 37 °C for 16 h. A single colony was used to inoculate 10 ml of LB media containing 50 μg $mL^{-1}$ kanamycin, and the cultures were incubated at 37 °C for 16 h. This starter culture was used to inoculate 600 ml of S-broth (35 g tryptone, 20 g yeast extract, 5 g NaCl, pH 7.4) containing 50 μg $mL^{-1}$ kanamycin, which was incubated with shaking (250 rpm) at 37 °C until it reached an $A_{600}$ of 0.7. After cooling to room temperature, IPTG was added to a final concentration of 0.4 mM, and incubation with shaking (200 rpm) continued at 18 °C for 16 h. Cells were harvested by centrifugation at 8000*g* for 20 min at 4 °C and then resuspended in 40 ml binding buffer (50 mM $NaP_i$, 300 mM NaCl, 5 mM imidazole, pH 7.5) containing protease inhibitor (Roche cOmplete EDTA-free protease inhibitor mixture) and lysozyme (0.1 mg $mL^{-1}$) by nutating at 4 °C for 30 min. Benzonase (1 μl, 250 U) was added to the mixture and then lysis was effected by sonication [10 × (15 s on/45 s off) at 45% amplitude]. The lysate was centrifuged at 18,000×*g* for 20 min at 4 °C and the supernatant was collected. The supernatants were filtered (0.45 μm) and loaded onto a 1 ml HisTrap column (GE). The column was washed with 3 × 10 ml of binding buffer, and the protein was eluted using elution buffer (50 mM $NaP_i$, 300 mM NaCl, 400 mM imidazole, pH 7.5). Fractions containing product, as judged by SDS-PAGE, were further purified by size-exclusion chromatography on a HiPrep 16/60 Sephacryl S-200 HR column (GE) using 50 mM $NaP_i$, 150 mM NaCl, pH 7.5. Protein concentration was determined using the bicinchoninic acid assay.

### *Enzyme assays for* Ra*GH31*

Hydrolytic activity was examined towards a range of 4-nitrophenyl glycosides (each 4 mM) in 600 μl reaction mixtures containing 50 nM *Ra*GH31 in 50 mM phosphate buffer, 150 mM NaCl, pH 7 at room temperature. Reactions were performed in a cuvette and monitored for release of PNP using an UV/Vis spectrophotometer at the isosbestic point of 4-nitrophenol (λ = 348 nm) where the extinction coefficient was 5125 $M^{-1}$ $cm^{-1}$ under the assay conditions. The following sugars were examined: α-D-glucopyranoside (PNPGlc), 4-nitrophenyl α-D-sulfoquinovoside (PNPSQ), 4-nitrophenyl α-D-galactopyranoside (PNPGal), 4-nitrophenyl α-D-xylopyranoside (PNPXyl), 4-nitrophenyl α-D-mannopyranoside (PNPMan), 4-nitrophenyl *N*-acetyl-α-D-galactosaminide (PNPGalNAc), 4-nitrophenyl α-D-glucuronide (PNPGlcA), and 4-nitrophenyl α-D-glucopyranoside 6-phosphate (6-P-PNPGlc).

Hydrolytic activities toward various disaccharides were assessed in 600 μl reaction mixtures containing 73.8 nM *Ra*GH31 in 50 mM phosphate buffer, 150 mM NaCl, pH 7 at room temperature. After 5 h, reactions were heat inactivated and then applied to an HP-TLC plate and eluted with ethyl acetate, methanol, and water (7:4:2). Products were visualized by spraying with a solution of 0.2% orcinol, 10% $H_2SO_4$, and 10% water in 80% ethanol and heating. The following sugars were assessed as substrates: trehalose [α-D-Glc*p*-(1↔1)-α-D-Glc*p*], kojibiose [α-D-Glc*p*-(1→2)-D-Glc], nigerose [α-D-Glc*p*-(1→3)-D-Glc], maltose [α-D-Glc*p*-(1→4)-D-Glc], isomaltose [α-D-Glc*p*-(1→6)-D-Glc], and sucrose [β-D-Fru*f*-(2↔1)-α-D-Glc*p*].

To establish the stereochemistry of the initially formed product of *Ra*GH31 catalyzed hydrolysis of PNPGlc, *Ra*GH31 (73.8 nM) was added to a solution of PNPGlc (10 mM) in 50 mM phosphate buffer, 150 mM NaCl at pD 7.4. $^1$H NMR (500 MHz) revealed the formation of a new signal at δ 5.22 ppm, $J_{1,2}$ 3.8 Hz, assigned α-glucose.

The Michaelis parameter $k_{cat}/K_M$ was measured for PNPGlc hydrolysis using the substrate depletion method in 50 mM phosphate buffer, 150 mM NaCl at a range of pH values (4.07, 5.09, 5.49, 5.98, 6.47, 7.01, 7.47, 8.04, 8.52, 9.29) at room temperature. A concentration of PNPGlc of 0.02 mM was used, being $<K_M/10$. Reactions were initiated by adding *Ra*GH31 to a final concentration of 50 nM to PNPGlc (0.02 mM) in buffer and the rate measured continuously using a UV/visible spectrophotometer. $k_{cat}/K_M$ values were calculated using the equation $y = (y_0 - y_\infty) \times \exp(-k \times t) + y_\infty$, where $k_{cat}/K_M = k/[E]$; $pK_a$ values were calculated using the Prism 5 software package (Graphpad Scientific Software) using the equation $y = m \times (1/(1+[(10^{-x})/(10^{-pKa1})) + (10^{-pKa2})/(10^{-x})]))+c$. Data for each pH were fitted to one phase decay curves to get $k_{cat}/K_M$ values. The data for $k_{cat}/K_M$ *versus* pH was fit to a bell-shaped curve, with an optimum at pH 6 (Fig. 2*A*). $pK_{a1}$ and $pK_{a2}$ were calculated as 5.90 ± 0.38 and 5.73 ± 0.38. Below pH 4.0 and above pH 9.5, the enzyme was unstable.

Temperature stability of *Ra*GH31 was assessed by incubation of 50 nM *Ra*GH31 in the assay buffer for 3 h at different temperatures (room temperature, 30 °C, 35 °C, 40 °C, 45 °C, 50 °C, 55 °C). After this time, PNPGlc was added to a final concentration of 0.02 mM, and the reaction rate measured using a UV/Vis spectrometer. After 3 h incubated at 30 °C, the enzyme only has 7% activity compared to room temperature. When the temperature raised to 40 °C, the enzyme only has 0.7% activity remaining.

*Michaelis–Menten kinetics* were measured for RaGH31-catalyzed hydrolysis of PNPGlc using a UV/visible spectrophotometer. Reactions were conducted in 50 mM sodium phosphate, 150 mM NaCl (pH 6) at 25 °C using 50 nM *Ra*GH31 at substrate concentrations ranging from 0.05 to 2 mM. For quantitative analysis of *Ra*GH31 hydrolysis of disaccharides, the reaction was measured by using Colorimetric Detection Kit (Invitrogen by Thermo Fisher Scientific). The reaction was performed by using 12.3 μM *Ra*GH31 in 50 mM sodium phosphate, 150 mM NaCl buffer (pH 6) with various concentrations of the disaccharide substrates. Samples at each concentration were measured for 10 and 20 min to ensure a linear initial rate for the reaction. The reaction was quenched by heating at 80 °C for 5 min. The quenched samples were cooled to room temperature and then glucose was released according to the manufacturer's instructions at 30 °C. The absorption was measured using a UV/visible spectrophotometer. Hydrolysis of maltose, nigerose, and trehalose

produce two equivalents of glucose, and the reaction rates were therefore halved. Kinetic parameters ($k_{cat}$, $K_M$, $k_{cat}/K_M$) were calculated using the Prism 5 software package (GraphPad Scientific Software) using the Michaelis–Menten equation.

## Data availability

The GH31 subfamily classification is publicly available on the CAZY database (http://www.cazy.org).

*Supporting information*—This article contains supporting information (39, 40, 60, 61).

*Author contributions*—N. T. and S. J. W. conceptualization; B. H. and N. T. methodology; T. A., B. H., and N. T. data curation; T. A., B. H., and N. T. formal analysis; J. L., N. M. S., E. D. G.-B., and S. J. W. investigation; N. T. validation; T. A., N. T., and S. J. W. writing–review and editing.

*Funding and additional information*—S. J. W. is supported by the Australian Research Council (DP210100233, DP210100235). E. D. G.-B. acknowledges support from The Walter and Eliza Hall Institute of Medical Research, National Health and Medical Research Council of Australia (NHMRC) project grant GNT2000517, the Australian Cancer Research Fund, and the Brian M. Davis Charitable Foundation Centenary Fellowship.

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

*Abbreviations*—The abbreviations used are: CAZy, Carbohydrate Active enzyme; CBM, carbohydrate-binding module; EC, enzyme commission; GH, glycoside hydrolase (family); 6GT, 6-α-glucosyltransferase; HMM, hidden Markov model; PDB, protein databank; SI, sucrase-isomaltase; SSN, sequence similarity network.

## References

1. Drula, E., Garron, M. L., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022) The carbohydrate-active enzyme database: functions and literature. *Nucl. Acids Res.* **50**, D571–d577
2. The CAZypedia Consortium (2018) Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. *Glycobiology* **28**, 3–8
3. Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, H., 3rd, and Henrissat, B. (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* **12**, 186
4. Stam, M. R., Danchin, E. G., Rancurel, C., Coutinho, P. M., and Henrissat, B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng. Des. Sel.* **19**, 555–562
5. Li, X., Kouzounis, D., Kabel, M. A., de Vries, R. P., and Dilokpimol, A. (2022) Glycoside hydrolase family 30 harbors fungal subfamilies with distinct polysaccharide specificities. *New Biotechnol.* **67**, 32–41
6. Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., and Henrissat, B. (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J.* **432**, 437–444
7. Mathieu, S., Henrissat, B., Labre, F., Skjåk-Bræk, G., and Helbert, W. (2016) Functional exploration of the polysaccharide lyase family PL6. *PLoS One* **11**, e0159415
8. Jongkees, S. A., and Withers, S. G. (2014) Unusual enzymatic glycoside cleavage mechanisms. *Acc. Chem. Res.* **47**, 226–235
9. Garron, M. L., and Cygler, M. (2010) Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. *Glycobiology* **20**, 1547–1573
10. Viborg, A. H., Terrapon, N., Lombard, V., Michel, G., Czjzek, M., Henrissat, B., *et al.* (2019) A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (GH16). *J. Biol. Chem.* **294**, 15973–15986
11. Mewis, K., Lenfant, N., Lombard, V., and Henrissat, B. (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization. *Appl. Environ. Microbiol.* **82**, 1686–1692
12. Kashiwabara, S., Azuma, S., Tsuduki, M., and Suzuki, Y. (2000) The primary structure of the subunit in *Bacillus thermoamyloliquefaciens* KP1071 molecular weight 540,000 homohexameric α-glucosidase II belonging to the glycosyl hydrolase family 31. *Biosci. Biotechnol. Biochem.* **64**, 1379–1393
13. Yamamoto, K., and Davis, B. G. (2012) Creation of an α-mannosynthase from a broad glycosidase scaffold. *Angew. Chem. Int. Ed. Engl.* **51**, 7449–7453
14. Ernst, H. A., Lo Leggio, L., Willemoës, M., Leonard, G., Blum, P., and Larsen, S. (2006) Structure of the *Sulfolobus solfataricus* α-glucosidase: implications for domain conservation and substrate recognition in GH31. *J. Mol. Biol.* **358**, 1106–1124
15. Kato, N., Suyama, S., Shirokane, M., Kato, M., Kobayashi, T., and Tsukagoshi, N. (2002) Novel α-glucosidase from *Aspergillus nidulans* with strong transglycosylation activity. *Appl. Environ. Microbiol.* **68**, 1250–1256
16. Crombie, H. J., Chengappa, S., Jarman, C., Sidebottom, C., and Reid, J. S. (2002) Molecular characterisation of a xyloglucan oligosaccharide-acting α-D-xylosidase from nasturtium (*Tropaeolum majus* L.) cotyledons that resembles plant 'apoplastic' alpha-D-glucosidases. *Planta* **214**, 406–413
17. Speciale, G., Jin, Y., Davies, G. J., Williams, S. J., and Goddard-Borger, E. D. (2016) YihQ is a sulfoquinovosidase that cleaves sulfoquinovosyl diacylglyceride sulfolipids. *Nat. Chem. Biol.* **12**, 215–217
18. Rozeboom, H. J., Yu, S., Madrid, S., Kalk, K. H., Zhang, R., and Dijkstra, B. W. (2013) Crystal structure of α-1,4-glucan lyase, a unique glycoside hydrolase family member with a novel catalytic mechanism. *J. Biol. Chem.* **288**, 26764–26774
19. Rahfeld, P., Wardman, J. F., Mehr, K., Huff, D., Morgan-Lang, C., Chen, H. M., *et al.* (2019) Prospecting for microbial α-*N*-acetylgalactosaminidases yields a new class of GH31 O-glycanase. *J. Biol. Chem.* **294**, 16400–16415
20. Zechel, D. L., and Withers, S. G. (2000) Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33**, 11–18
21. Lee, S. S., Yu, S., and Withers, S. G. (2003) Detailed dissection of a new mechanism for glycoside cleavage: α-1,4-glucan lyase. *Biochemistry* **42**, 13081–13090
22. Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313
23. Yamamoto, T., Unno, T., Watanabe, Y., Yamamoto, M., Okuyama, M., Mori, H., *et al.* (2004) Purification and characterization of *Acremonium implicatum* α-glucosidase having regioselectivity for alpha-1,3-glucosidic linkage. *Biochim. Biophys. Acta* **1700**, 189–198
24. Sampedro, J., Sieiro, C., Revilla, G., González-Villa, T., and Zarra, I. (2001) Cloning and expression pattern of a gene encoding an α-xylosidase active against xyloglucan oligosaccharides from Arabidopsis. *Plant Physiol.* **126**, 910–920
25. Larsbrink, J., Izumi, A., Hemsworth, G. R., Davies, G. J., and Brumer, H. (2012) Structural enzymology of *Cellvibrio japonicus* Agd31B protein reveals α-transglucosylase activity in glycoside hydrolase family 31. *J. Biol. Chem.* **287**, 43288–43299
26. Sim, L., Willemsma, C., Mohan, S., Naim, H. Y., Pinto, B. M., and Rose, D. R. (2010) Structural basis for substrate selectivity in human maltase-glucoamylase and sucrase-isomaltase N-terminal domains. *J. Biol. Chem.* **285**, 17763–17770
27. Gray, G. M., Lally, B. C., and Conklin, K. A. (1979) Action of intestinal sucrase-isomaltase and its free monomers on an alpha-limit dextrin. *J. Biol. Chem.* **254**, 6038–6043

28. Ren, L., Qin, X., Cao, X., Wang, L., Bai, F., Bai, G., *et al.* (2011) Structural insight into substrate specificity of human intestinal maltase-glucoamylase. *Protein Cell* **2**, 827–836
29. Roig-Zamboni, V., Cobucci-Ponzano, B., Iacono, R., Ferrara, M. C., Germany, S., Bourne, Y., *et al.* (2017) Structure of human lysosomal acid α-glucosidase-a guide for the treatment of Pompe disease. *Nat. Commun.* **8**, 1111
30. Chaudet, M. M., and Rose, D. R. (2016) Suggested alternative starch utilization system from the human gut bacterium *Bacteroides thetaiotaomicron*. *Biochem. Cell Biol.* **94**, 241–246
31. Lovering, A. L., Lee, S. S., Kim, Y. W., Withers, S. G., and Strynadka, N. C. (2005) Mechanistic and structural analysis of a family 31 alpha-glycosidase and its glycosyl-enzyme intermediate. *J. Biol. Chem.* **280**, 2105–2115
32. Okuyama, M., Kaneko, A., Mori, H., Chiba, S., and Kimura, A. (2006) Structural elements to convert Escherichia coli alpha-xylosidase (YicI) into alpha-glucosidase. *FEBS Lett.* **580**, 2707–2711
33. Larsbrink, J., Izumi, A., Ibatullin, F. M., Nakhai, A., Gilbert, H. J., Davies, G. J., and Brumer, H. (2011) Structural and enzymatic characterization of a glycoside hydrolase family α-xylosidase from *Cellvibrio japonicus* involved in xyloglucan saccharification. *Biochem. J.* **436**, 567–580
34. Cao, H., Walton, J. D., Brumm, P., and Phillips, G. N., Jr. (2020) Crystal structure of α-xylosidase from *Aspergillus niger* in complex with a hydrolyzed xyloglucan product and new insights in accurately predicting substrate specificities of GH31 family glycosidases. *ACS Sustain. Chem. Eng.* **8**, 2540–2547
35. Aga, H., Maruta, K., Yamamoto, T., Kubota, M., Fukuda, S., Kurimoto, M., *et al.* (2002) Cloning and sequencing of the genes encoding cyclic tetrasaccharide-synthesizing enzymes from *Bacillus globisporus* C11. *Biosci. Biotechnol. Biochem.* **66**, 1057–1068
36. Light, S. H., Cahoon, L. A., Halavaty, A. S., Freitag, N. E., and Anderson, W. F. (2016) Structure to function of an α-glucan metabolic pathway that promotes *Listeria monocytogenes* pathogenesis. *Nat. Microbiol.* **2**, 16202
37. Light, S. H., Cahoon, L. A., Mahasenan, K. V., Lee, M., Boggess, B., Halavaty, A. S., *et al.* (2017) Transferase *versus* hydrolase: the role of conformational flexibility in reaction specificity. *Structure* **25**, 295–304
38. Fujimoto, Z., Suzuki, N., Kishine, N., Ichinose, H., Momma, M., Kimura, A., *et al.* (2017) Carbohydrate-binding architecture of the multi-modular α-1,6-glucosyltransferase from *Paenibacillus* sp. 598K, which produces α-1,6-glucosyl-α-glucosaccharides from starch. *Biochem. J.* **474**, 2763–2778
39. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589
40. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucl. Acids Res.* **50**, D439–D444
41. Mukai, K., Maruta, K., Satouchi, K., Kubota, M., Fukuda, S., Kurimoto, M., *et al.* (2004) Cyclic tetrasaccharide-synthesizing enzymes from *Arthrobacter globiformis* A19. *Biosci. Biotechnol. Biochem.* **68**, 2529–2540
42. Aga, H., Nishimoto, T., Kuniyoshi, M., Maruta, K., Yamashita, H., Higashiyama, T., *et al.* (2003) 6-Alpha-glucosyltransferase and 3-alpha-isomaltosyltransferase from *Bacillus globisporus* N75. *J. Biosci. Bioeng.* **95**, 215–224
43. Goddard-Borger, E. D., and Williams, S. J. (2017) Sulfoquinovose in the biosphere: occurrence, metabolism and functions. *Biochem. J.* **474**, 827–849
44. Abayakoon, P., Jin, Y., Lingford, J. P., Petricevic, M., John, A., Ryan, E., *et al.* (2018) Structural and biochemical insights into the function and evolution of sulfoquinovosidases. *ACS Cent. Sci.* **4**, 1266–1273
45. Miyazaki, T., Ishizaki, Y., Ichikawa, M., Nishikawa, A., and Tonozuka, T. (2015) Structural and biochemical characterization of novel bacterial α-galactosidases belonging to glycoside hydrolase family 31. *Biochem. J.* **469**, 145–158
46. Ikegaya, M., Moriya, T., Adachi, N., Kawasaki, M., Park, E. Y., and Miyazaki, T. (2022) Structural basis of the strict specificity of a bacterial GH31 α-1,3-glucosidase for nigerooligosaccharides. *J. Biol. Chem.* **298**, 101827
47. Miyazaki, T., and Park, E. Y. (2020) Crystal structure of the *Enterococcus faecalis* α-N-acetylgalactosaminidase, a member of the glycoside hydrolase family 31. *FEBS Lett.* **594**, 2282–2293
48. Valk, V., Kaaij, R. M.v. d., and Dijkhuizen, L. (2017) The evolutionary origin and possible functional roles of FNIII domains in two Microbacterium aurum B8.A granular starch degrading enzymes, and in other carbohydrate acting enzymes. *Amylase* **1**, 1–11
49. Helbert, W., Poulet, L., Drouillard, S., Mathieu, S., Loiodice, M., Couturier, M., *et al.* (2019) Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6063–6068
50. Meek, R. W., Brockerman, J., Fordwour, O. B., Zandberg, W. F., Davies, G. J., and Vocadlo, D. J. (2022) The primary familial brain calcification-associated protein MYORG is an α-galactosidase with restricted substrate specificity. *PLoS Biol.* **20**, e3001764
51. Yang, F., Abdelnabby, H., and Xiao, Y. (2015) A mutant of the nematophagous fungus *Paecilomyces lilacinus* (Thom) is a novel biocontrol agent for Sclerotinia sclerotiorum. *Microb. Pathog.* **89**, 169–176
52. Vocadlo, D. J., Wicki, J., Rupitz, K., and Withers, S. G. (2002) A case for reverse protonation: identification of Glu160 as an acid/base catalyst in *Thermoanaerobacterium saccharolyticum* beta-xylosidase and detailed kinetic analysis of a site-directed mutant. *Biochemistry* **41**, 9736–9746
53. Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4**, e4345
54. Jones, D. R., Thomas, D., Alger, N., Ghavidel, A., Inglis, G. D., and Abbott, D. W. (2018) Saccharis: an automated pipeline to streamline discovery of carbohydrate active enzyme activities within polyspecific families and *de novo* sequence datasets. *Biotechnol. Biofuels* **11**, 27
55. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402
56. Li, W., and Godzik, A. (2006) Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659
57. Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780
58. Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucl. Acids Res.* **39**, W29–37
59. Letunic, I., and Bork, P. (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucl. Acids Res.* **47**, W256–W259
60. Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., *et al.* (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–10254
61. Satoh, T., Toshimori, T., Yan, G., Yamaguchi, T., and Kato, K. (2016) Structural basis for two-step glucose trimming by glucosidase II involved in ER glycoprotein quality control. *Sci. Rep.* **6**, 20575