**Divergent evolution for diverse substrate recognition by
family 31 glycoside hydrolases**

Journal:	<i>Biochemistry and Cell Biology</i>
Manuscript ID	bcb-2016-0022.R2
Manuscript Type:	Article
Date Submitted by the Author:	06-Apr-2016
Complete List of Authors:	Chaudet, Marcia; University of Waterloo, Rose, David; University of Waterloo, Biology
Keyword:	Glycoside hydrolase, α -glucosidase, starch digestion, Family GH31, phylogenetic

Divergent evolution for diverse substrate recognition by family 31 glycoside hydrolases.

Chaudet, M.M.¹ and Rose, D.R.^{1*}

¹University of Waterloo, 200 University Ave West, Waterloo, ON, Canada, N2L 3G1.

***Correspondence to: David R. Rose, Department of Biology, University of Waterloo, 200 University Ave. West, Waterloo, ON, Canada, N2L 3G1, email: drrose@uwaterloo.ca**

Draft

Abstract

Carbohydrates make up an important component of our diet contributing a significant portion to our total dietary caloric intake. The ability to harvest these molecules for energy is reliant on the mechanisms of carbohydrate-active enzymes.

Family 31 α -glucosidases are a group of glycoside hydrolases that has been shown to play a key role in the metabolic process of hydrolyzing dietary starch into the monomer of glucose. The purpose of the research presented here is to explore evolutionary changes that occurred within this **family** of glycoside hydrolases and to relate these divergences to observed structural differences in relation to predicted substrate preferences. Here we report specific single amino acid changes that are believed to **have** arisen during evolution and are directly related to the ability of these enzymes to bind different starch-based glycans. Through phylogenetic analysis we are able to observe a number of evolutionary adaptations that we believe resulted in duplicated genes that allow for the efficient utilization of dietary starch.

Keywords

Glycoside hydrolase, α -glucosidase, starch digestion, phylogenetic, **family** GH31

Introduction

Carbohydrates play an important role in providing nutritional components from our diet. The physiochemical properties of carbohydrates can be quite diverse in relation to the simple sugars that make up their structure as well as their specific glycosidic linkages. As a result of their diversity, humans are dependent on enzymatic means to utilize and harvest these molecules for energy. Glycoside hydrolases are a group of enzymes that have the ability to break the glycosidic bond between two sugar residues (Davies and Henrissat 1995). These enzymes demonstrate a range of substrate preferences and as a result are classified into discrete families known as carbohydrate-active enzyme families (Cantarel et al. 2009). These families demonstrate an overall conserved architecture and have

conserved domains highlighting their evolutionary stability(Davies and Henrissat 1995).

The substrate specificity of some glycoside hydrolases is very precise, whereas others can be permissive, with the ability to act on linkages with a range of stereochemistry or sugar substituents(Henrissat and Davies 1997). Substrate specificity can change within an enzyme family in order to accommodate environmental pressure from substrate exposure resulting in evolutionary changes. These changes can arise from gene duplication within an organism allowing for multiple genes to perform similar functions. When such an event occurs it can result in discrete changes, which ultimately can influence the secondary structure creating new functions within a single enzyme family(Koonin 2005).

Family GH31 is a group of glycoside hydrolases found in all domains of life and can be observed in a variety of environments from acidic hotsprings to the human digestive system(Rolfsmeier and Blum 1995, Nichols et al. 1998). This family of enzymes encompasses a range of diverse activities that include α -glucosidase (EC 3.2.1.20); α -galactosidase (EC 3.2.1.22); α -mannosidase (EC 3.2.1.24); α -1,3-glucosidase (EC 3.2.1.84); sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); α -xylosidase (EC 3.2.1.177); α -glucan lyase (EC 4.2.2.13); isomaltosyltransferase (EC 2.4.1.-) and oligosaccharide α -1,4-glucosyltransferase (EC 2.4.1.161). Family GH31 are of interest in relation to the human diet, specifically α -glucosidases, which are important enzymes in primary metabolism. α -glucosidases are responsible for the hydrolysis of dietary starch molecules resulting in the release of free-glucose. In humans, two family GH31 α -glucosidases are known to play a key role in the final stages of starch digestion and are identified as maltase glucoamylase (MGAM) and sucrose-isomaltase (SI)(Hunziker et al. 1986, Van Beers et al. 1995, Nichols et al. 2003, 2009, Quezada-Calvillo et al. 2008). These enzymes each contain two independently functional domains (N-terminal and C-terminal) and are believed to have evolved from a single ancestral gene(Nichols et al. 2003, Naumoff 2007). The environmental pressure that has led to duplication and divergence is believed to be

related to the structural complexity of its target substrate, starch(Hunziker et al. 1986, Lin et al. 2014, Pontremolil et al. 2015).

A number of family 31 α -glucosidase structures have been solved, providing important structural information that can be related to function. These models include one from the Archaea domain, an α -glucosidase from *Sulfolobus solfataricus* (MalA, PDB code: 2G3M)(Ernst et al. 2006). Three bacterial α -glucosidases that come from organisms inhabiting the human gut; two from the Gram-negative bacterium *Bacteroides thetaiotaomicron* (BT_3299, PDB code: 5DJW; BT_0339, PDB code: 5F7C) and one from the Gram-positive bacterium *Ruminococcus obeum* (renamed *Blausti obeum*)(Ro- α G1, PDB code: 3MKK)(Tan et al. 2010, Chaudet and Rose 2016). Four structures from the eukaryote domain that include α -glucosidases from *Homo sapiens* (NT_MGAM, PDB code: 2QMJ; CT_MGAM, PDB code: 3TOP; NT_SI, PDB code: 3LPP) and an α -glucosidase from the plant species *Beta vulgaris* (SBG, PDB code: 3W37)(Sim et al. 2008, 2010, Ren et al. 2011, Tagami et al. 2013). Overall, these structures share a conserved three-dimensional architectural homology characterized as having four domains and 2 subdomains: an N-terminal β -sandwich domain, $(\beta/\alpha)_8$ barrel domain, insertion subdomain 1 and 2, and proximal C-terminal domain(Janeček et al. 2007, Tagami et al. 2013).

Family 31 α -glucosidases demonstrate conserved catalytic residues with an aspartic acid as the nucleophile and a separate aspartic acid acting as an acid/base catalyst, which operate through a retaining mechanism of catalysis. Despite these enzymes having a common role of hydrolyzing starch glycans, it has been observed that these enzymes demonstrate varied activities(Sim et al. 2010, Lee et al. 2012, Chaudet et al. 2012, Lin et al. 2014). The range of their ability includes hydrolyzing α , 1-4 and/or α , 1-6 linkages between glucose residues or hydrolyzing short and/or long linear maltooligosaccharides(Tan et al. 2010, Lee et al. 2012, Chaudet et al. 2012, Tagami et al. 2013)(Table 1). The ability of this family to accommodate a range of starch

glycans is believed to result from the variations in starch structure that these enzymes are exposed to in their selective environments.

In this current study we are interested in investigating specific evolutionary changes that may have occurred over time that influence their ability to utilize starch substrates and ultimately result in changes in the secondary structure. Here we investigate the phylogenetic evolution of structurally solved **family 31** α -glucosidases and assess the phylogenetic relationship within this family. We assess structural differences that have evolved within **family** GH31 α -glucosidases and provide evidence of single amino acid changes that contribute to functional divergences.

Methods

*Phylogenetic Analysis of **Family 31** α -glucosidase Structures*

Table 1 indicates the structures that were investigated in this study. We focused our analysis on experimentally solved structures of α -glucosidases as part of **family 31** glycoside hydrolases. The computer program MUSTANG was used to generate a structure based multiple sequence alignment by aligning the protein structures (Konagurthu et al. 2010). FastTree was used to generate a maximum-likelihood phylogenetic tree using the structure sequence alignment from MUSTANG (Price et al. 2009, 2010). Geneious software program was used to analyze alignments and phylogenetic trees (Kearse et al. 2012). Constructed trees were unrooted, as no assumption was made regarding which sequence is closest to the common ancestor.

*Predicted Evolutionary Changes in **Family 31** α -glucosidase Structures*

Phylogenetic analysis was completed in the program DIVERGE by the neighbor-joining method using the structure sequence alignment and phylogenetic tree (Gu and Vander Velden 2002, Gu et al. 2013). Analysis of predicted changes were then mapped to the Protein Data Bank files for the solved structures (Berman et al. 2000). These evolutionary differences were then assessed manually with a focus on regions

related to catalysis and substrate binding (Table 2). Analysis of these predicted residues was done with care due to the limitations of the sample population of 8 structures as reported in the PDB.

Comparison of Predicted Evolutionary Changes and Acarbose Binding

To aid in the analysis of predicted evolutionary changes, structures were analyzed in relation to a bound acarbose molecule. Structures that were experimentally solved with a bound acarbose included NT_MGAM (PDB code: 2QMJ), CT_MGAM (PDB code: 3TOP) and SBG (PDB code: 3W37)(Sim et al. 2008, Ren et al. 2011, Tagami et al. 2013). The structures of BT_3299 (PDB code: 5DJW) and BT_0339 (PDB code: 5F7C) were previously reported with a modeled acarbose molecule(Chaudet and Rose 2016). Molecular modeling of acarbose was completed using Coot for structures that were not experimentally solved with a bound acarbose(Emsley and Cowtan 2004). The structure of Ro- α G1 (PDB code: 3MKK) was previously solved with a bound isomaltose molecule, which was used as comparison to guide the placement of an acarbose molecule(Tan et al. 2010). The homologous structure of NT_MGAM (PDB: 2QMJ) was also used in comparison with Ro- α G1. The structure of NT_SI (PDB code: 3LPP) was previously solved with a bound inhibitor kotalanol, which was used as comparison to guide the placement of an acarbose molecule(Sim et al. 2010). The structure 2QMJ was used to guide the placement of any remaining acarbose rings that are absent in kotalanol. Acarbose was modeled into the structure of MalA (PDB code: 2G3M) by using the structure NT_MGAM as guidance. In all cases the acarbose molecule was modeled in by superimposing the substrate onto the bound molecules in the comparative structures and used as a guide to place the inhibitor molecule.

Phylogenetic Analysis of Characterized Family 31 α -glucosidases

Sequences used in the evolutionary analysis of characterized family 31 α -glucosidases are shown in Supplementary Table 1. Selected sequences include those that have been characterized as α -glucosidases and have the designated enzyme

commission number of 3.2.1.20. The computer program MUSCLE was used to generate the multiple sequence alignment(Edgar 2004). Gaps were removed prior to phylogenetic tree construction to ensure accurate evolutionary analysis. FastTree was used to generate a maximum-likelihood phylogenetic tree using the modified multiple sequence alignment(Price et al. 2009, 2010). As referred to above, the software program Geneious was used to analyze alignments and phylogenetic trees(Kearse et al. 2012) with the constructed tree being un-rooted, due to no assumption regarding which sequence is closest to the common ancestor.

Results

Structural Evolution

The first step in the evolutionary analysis was the generation of a structural multiple sequence alignment of the structures as experimentally determined for [family](#) 31 α -glucosidases (Supplementary Figure 1). Overall, there is little sequence similarity with some regions demonstrating high conservation. One of the common reported features of [family](#) 31 α -glucosidases is their signature catalytic sequence WiDMNE, where 'i' is variable(Ernst et al. 2006). Observing this sequence in the sequence alignment we see there is a large variation in this region in [B. thetaiotaomicron](#) BT_0339 (highlighted in Supplementary Figure 1). These differences were also observed structurally and were reported to reflect differences in substrate preferences(Chaudet and Rose 2016). This initially gives us some insight into possible evolutionary differences within this [family](#) and interestingly within [B. thetaiotaomicron](#).

A phylogenetic tree was constructed by the maximum-likelihood method using the above alignment (Figure 1). Overall, the branch pattern in this tree was not surprising with eukaryotic and prokaryotic clusters. Interestingly, [B. thetaiotaomicron](#) BT_3299 and BT_0339 are not evolutionarily close in this phylogenetic tree with BT_0339 showing close relatedness to the archeal organism [S. solfataricus](#). Based on this observation, there appears to be some evolutionary

change that occurred within this organism in relation to the mechanism of an α -glucosidic enzyme. It also worth noting that *R. obeum* appears to cluster independently and is can be observed that this difference has arisen from the discrete substrate preference of this α -glucosidase *for* isomaltose.

Variability in Catalytic Region

The program DIVERGE *was used to analyze functional divergence for structurally solved family GH31 enzymes α -glucosidases. By using the provided structure based multiple sequence alignment and the given phylogenetic tree, DIVERGE detected site-specific changes in evolutionary rate*(Gu and Vander Velden 2002). *Analyzing these predictions in relation to biological evidence* resulted in two *structural regions where functional divergences are predicted to have occurred. These regions include* the catalytic site and a looped substrate-binding region (Table 2). Observing the first predicted change that occurs in the catalytic site allowed for the observation that this change could be involved *in* binding substrates in the +1 and +2 subsites. When we assess this amino acid change individually in each structure we find support of this residue having a role in substrate binding (Figure 2). The three human structures all demonstrate residues that interact with bound acarbose as observed with Thr204, Gln1158 and Gln232 in NT_MGAM, CT_MGAM and NT_SI respectively. The α -glucosidase structure from *B. vulgaris* was observed to have a predicted change to the amino acid Ile233 and structurally is in close proximity to the bound acarbose inhibitor in the +2 subsite. The structure *of* MalA has the predicted amino acid Asp346 observed to closely interact with a modeled acarbose in the +1 sub-site. Ro- α G1 structure was observed to have a predicted to change to the residue Asp74, *which* directly interacts with the bound acarbose substrate in the +1 subsite. Both *B. thetaiotaomicron* structures demonstrate a predicted amino acid change that is believed to influence substrate binding as observed through residues Tyr91 and Gly229 for BT_3299 and BT_0339 respectively. Tyr91 in the BT_3299 model is specifically believed to be involved in stabilizing bound substrates in the active site in the +1 subsite.

Variability in Looped Substrate-Binding Region

The second predicted evolutionary change as observed in DIVERGE occurs in a looped substrate-binding region in the assessed structures. The observed single amino acid change resides in the second subdomain just outside of the active site and is proposed to be involved in stabilizing longer bound substrates (Tagami et al. 2013). Predicted evolutionary changes in MalA and Ro- α G1 (*B. thetaiotaomicron* structures has unresolved regions that were not visible) are observed to occur in a helix bundle that is explicitly observed in these prokaryote structures. These predicted changes include Arg334 and Asp333 in MalA and Ro- α G1 respectively (Figure 3A).

Predicted changes in the looped binding region were also observed in the eukaryote structures and are positioned to also play a key role in substrate binding. Predicted changes include Arg471, His1449, Asp500 and Asn493 for NT_MGAM, CT_MGAM, NT_SI and SBG respectively (Figure 3A). All of these changes reside in a loop that is observed to be absent in the four prokaryote structures (Figure 3B). This observation gives insight into the possibility that this region evolved to accommodate specific starch glycans that are uniquely encountered by eukaryote α -glucosidases.

Phylogenetic Relationships in Characterized *Family 31* α -glucosidases

A phylogenetic tree was constructed using characterized α -glucosidase protein sequences using the maximum-likelihood method (Figure 4). This analysis resulted in a number of branch clusters. The first cluster is composed of eukaryotic organisms consisting of mammals and plant species. The second eukaryotic cluster mainly includes fungal organisms. Archaea and bacteria form a cluster independently of the eukaryotic organisms. There are a number of interesting features of this tree. BT_3299 α -glucosidase from *B. thetaiotaomicron* is distant phylogenetically from BT_0339 and clusters more closely with the archaea group. Another interesting quality of this tree is the location of the *Homo sapiens* α -

glucosidases. NT_MGAM and NT_SI cluster together, where as CT_MGAM and CT_SI are more distant. It is observed that CT_SI demonstrates an evolutionary relatedness to *Oryctolagus cuniculus* (rabbit) and NT_MGAM shows relatedness to *Rattus norvegicus* (rat).

Eukaryote Signature Sequence

Comparison of the phylogenetic tree of characterized α -glucosidases with the solved structures allowed for the observation of an additional signature sequence within this [family](#) (Supplementary Figure 3A). The sequence occurs at positions 877-879 in the multiple sequence alignment and is identified as 'ppY, where p is variable but predominantly proline. Interestingly, it is observed that this sequence is conserved in only characterized eukaryotic α -glucosidases and is not present in characterized prokaryote α -glucosidases. This region of eukaryotic conservation happens to fall in the evolved loop in subdomain2 and is related to the second predicted evolutionary change that occurs in the looped binding region (Supplementary Figure 3B).

Discussion

Phylogenetic analysis derived from structural sequence alignment allowed for observations of [family](#) 31 α -glucosidases in relation to their ability to utilize maltooligosacchrides. The α -glucosidase from [R. obeum](#) is previously reported to preferentially hydrolyze isomaltose in comparison to [linear maltooligosacchrides](#) (Tan et al. 2010). This specificity is unique amongst the structurally analyzed [family](#) 31 α -glucosidases and we believe is reflected in its branching in the phylogenetic tree. This organism resides in the human colon, which is a diverse competitive environment. In order to out compete surrounding organisms perhaps [R. obeum](#) developed a distinct α -glucosidic ability to use starch-derived glycans in the human colon. The observation of BT_0339 having a higher relatedness to Mala [as observed in Figure 1 and 4](#) reveals that this enzyme may have some ancient origin and that BT_3299 is more closely related to the eukaryote organisms. As previously reported, NT_MGAM is more closely related to NT_SI and as evident in the tree, these enzymes cluster closer together than to CT_MGAM. Though these N-terminal domains have

slightly different substrate preferences, it can be observed that they both in general are able to bind and hydrolyze bulkier maltooligosacchrides (linear starch glycans and branched starch glycans) and this could reflect their evolutionary relatedness. This observation contributes to the hypothesis that these domains originate from a common ancestral gene(Nichols et al. 2003).

A number of changes that we propose influence substrate binding in family 31 α -glucosidases can be observed from this analysis. The first change involves residues in the +1 and +2 sugar binding site in the catalytic centre. This was observed in NT_MGAM with evidence from previous structural observations. It was previously observed that Asp203 was involved in binding acarbose in the +1 sub-site of NT_MGAM and provides evidence that a change at the amino acid Thr204 could influence substrate binding(Sim et al. 2008). The structure for NT_SI demonstrated a water molecule interacting with Gln232 similarly to what we are predicting here. A change to this specific residue would have specifically influenced sugars binding in the +1 subsite of NT_SI. Similarly, it was observed in CT_MGAM that D1157 interacted directly with acarbose in the +1 subsite and therefore provides support for our prediction that a change to the amino acid Gln1158 could influence the binding of starch molecules. Structural mutational studies of Ro- α G1 resulted in observed structural differences in the residue Asp73 resulting in observed structural disorder in loop β 7c α 7c and β 8c α 8c causing a lack of bound substrate in the active site(Tan et al. 2010). Enzyme assays involving this mutant (D73A) resulted in significant decrease in enzyme activity in the presence of maltooligosacchrides(Tan et al. 2010). These observations provide support for our proposal that the neighboring residue Asp74 has a key role in substrate binding and an evolutionary change to this residue would have influenced the catalytic abilities of this enzyme. It is also thought that a similar interaction exists within the *Bacteroides* structures based on modeling of acarbose within the models. Based on these results we suggest that a change in the +1 and +2 sugar binding site could strongly impact how substrates bind overall in the catalytic site of family 31 α -glucosidases.

The second predicted evolutionary change involves a secondary site that is proposed to be involved in substrate binding. Predicted changes to the helix bundle observed in MalA and Ro- α G1 are supported by structural observations. In the structure of Ro- α G1 researchers reported that the helix bundle was an important element in the dimerization of the crystal structure(Tan et al. 2010). This structural feature was also observed with MalA and it was reported that the helix bundle was an integral component to stabilizing the crystal packing(Ernst et al. 2006). These observations support our prediction that the evolutionary change of a specific residue in the helix bundle is believed to influence the secondary structure of these enzymes and ultimately could play a role in catalysis. A similar predicted change was observed within a loop specific to the eukaryotic structures and is also supported by previous structural observations. This structural feature was predicted to be involved in binding long starch glycans in the +4 subsite in SBG structure(Tagami et al. 2013). NT_MGAM, CT_MGAM and NT_SI are all observed to have the ability to utilize linear maltooligosacchrides and therefore, are structurally thought to involve additional binding sites outside the catalytic site of -1/+1 subsites (Figure 3B). CT_SI has been observed to utilize both sucrose and maltose-derived substrates; further supporting the evolved ability of this group of enzymes(Quezada-Calvillo et al. 2008). This predicted evolutionary divergence is further supported in the recognition of a eukaryotic specific signature sequence of 'ppY', as our results suggest that this characteristic region allows for the binding larger starch-based sugars. This gives support to the previous observation of the SBG structure and the prediction of additional sugar binding subsites existing in this region, which allows this plant α -glucosidase to bind and hydrolyze long-chain maltooligosaccharides(Tagami et al. 2013). We believe this evolved ability to accommodate longer starch substrates is reflected in the evolution of our dietary starch. Over evolution the availability of dietary starch has increased and as a result our digestive system has been exposed to an increase in diverse starch based polymers. This environmental change has allowed for our α -glucosidases (MGAM and SI) to adapt in order to extract the maximum amount of glucose from our diet.

In general, our phylogenetic analysis demonstrated typical clustering of branches amongst domains and phyla, but also provided evidence of interesting patterns of evolution. BT_0339 and BT_3299 from *B. thetaiotaomicron* were observed to not cluster together as one would imagine with BT_0339 being closer related to certain archeal organisms. Based on this result we believe these enzymes did not arise from a gene duplication event and that perhaps one of these paralogs was obtained by means of horizontal gene transfer. Alternatively, this divergence could reflect some form of species divergence that is specific for *B. thetaiotaomicron*. It is also believed that BT_0339 being closely related to genes derived from the archaea domain reflects the origin of BT_0339 being evolutionarily ancient as stated above. These paralogs in *B. thetaiotaomicron* could be evolution artifacts, gene redundancies that have yet to be removed from the genome, or alternatively they could perform discrete functions that are necessary for utilizing diverse forms of starch that exist in our diet. Focusing on the specific functions observed for the α -glucosidases involved in human digestion, we propose that these enzymes adapted alternative mechanisms in order to be able to utilize available substrates derived from dietary starch. The gut microbiome is a highly populated and competitive community and as a result inhabiting organisms must acquire or evolve enzymatic mechanisms in order to thrive in such a population. The phylogenetic relationship between the human enzymes MGAM and SI is one that is interesting, and our results support the notion that the divergence of these enzymes likely occurred prior to mammalian evolution as a result of the relatedness to distant species (Naumoff 2007). It is likely from our analysis that the common ancestor of these enzymes is one that occurs quite distantly in the phylogenetic tree of life in relation to *H. sapiens* and the sequence differences that exist between MGAM and SI likely occurred prior to the emergence of *H. sapiens*.

The observed evolutionary changes that have occurred within this family of enzymes are focused on changes that influence substrate binding. Overall, we propose that these changes influence the ability to utilize longer starch glycans in

addition to smaller oligosaccharides (Tagami et al. 2013). Particularly, the ability of eukaryote organisms to degrade long linear maltooligosaccharides and prokaryotes purposefully hydrolyzing smaller branched polymers. Changes involved in catalysis would likely result in alterations in conserved motifs that are signatures of family 31. However, changes to substrate binding are outside the core active site and would not influence the conserved overall secondary structure, as evident in the CAZy database. As a result of the evolutionary adaptations to the members of this family, there is an overall ability to utilize the polymer starch, but also there are observed discrete functions that reflect phylogenetic specific substrate preferences. Furthermore, this family of enzymes in the context of human digestion reveals an evolutionary symbiosis that developed over time, resulting in the mutual relationship of utilizing dietary starch. Understanding the evolutionary changes that occur within an enzyme family can help define structural studies and provide evidence for biochemical studies of substrate utilization within carbohydrate-active families. The results shown here provide brief insight into the phylogenetic relationship within this family and highlight the need for further analysis into the relationship between ancestor and descendants of family GH31 α -glucosidases.

Supplementary Material

Additional supplementary information can be found in the on-line version of this article.

Acknowledgments

The authors thank Dr. Laura Hug, University of Waterloo, for her guidance with the phylogenetic analysis. We would also like to thank Buford Nichols as part of the Starch Digestion Consortium for his insight into the relationship of these results to starch digestion. This work was supported by research grants from the National Science and Engineering Council of Canada (NSERC, 03740) and the Canadian Institutes of Health Research (CIHR, MOP 11237).

Reference

- Van Beers, E.H., Al, R.H., Rings, E.H., Einerhand, a W., Dekker, J., and Büller, H. a. 1995. Lactase and sucrase-isomaltase gene expression during Caco-2 cell differentiation. *Biochem. J.* **308** (Pt 3): 769–775.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**(1): 235–242.
- 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. *Nucleic Acids Res.* **37**(Database issue): D233–8.
- Chaudet, M.M., Allen, J.-L., and Rose, D.R. 2012. Expression and purification of two Family GH31 α -glucosidases from *Bacteroides thetaiotaomicron*. *Protein Expr. Purif.* **86**(2): 135–41.
- Chaudet, M.M., and Rose, D.R. 2016. Alternative starch utilization system from the human gut bacterium *Bacteroides thetaiotaomicron*. *Can. J. Biochem. Cell Biol.* doi:10.1139/bcb-2016-0002.
- Davies, G., and Henrissat, B. 1995. Structures and mechanisms of glycosyl hydrolases. *Structure* **3**(9): 853–9.
- Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Res.* **32**(5): 1792–1797.
- Emsley, P., and Cowtan, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D. Biol. Crystallogr.* **60**: 2126–32.
- 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.
- Gu, X., and Vander Velden, K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* **18**(3): 500–501.
- Gu, X., Zou, Y., Su, Z., Huang, W., Zhou, Z., Arendsee, Z., and Zeng, Y. 2013. An Update of DIVERGE Software for Functional Divergence Analysis of Protein Family. *Mol. Biol. Evol.* **30**(7): 1713–1719.
- Henrissat, B., and Davies, G. 1997. Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol.* **7**(5): 637–644.

- Hunziker, W., Spiess, M., Semenza, G., and Lodish, H.F. 1986. The sucrase-isomaltase complex: primary structure, membrane-orientation, and evolution of a stalked, intrinsic brush border protein. *Cell* **46**(2): 227–34.
- Janeček, Š., Svensson, B., and MacGregor, E.A. 2007. A remote but significant sequence homology between glycoside hydrolase clan GH-H and family GH31. *FEBS Lett.* **581**(7): 1261–1268.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**(12): 1647–9.
- Konagurthu, A.S., Reboul, C.F., Schmidberger, J.W., Irving, J. a., Lesk, A.M., Stuckey, P.J., Whisstock, J.C., and Buckle, A.M. 2010. MUSTANG-MR structural sieving server: Applications in protein structural analysis and crystallography. *PLoS One* **5**(4).
- Koonin, E. V. 2005. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **39**(1): 309–338.
- Lee, B.-H., Eskandari, R., Jones, K., Reddy, K.R., Quezada-Calvillo, R., Nichols, B.L., Rose, D.R., Hamaker, B.R., and Pinto, B.M. 2012. Modulation of starch digestion for slow glucose release through “toggling” of activities of mucosal α -glucosidases. *J. Biol. Chem.* **287**(38): 31929–38.
- Lin, A.H., Ao, Z., Quezada-calvillo, R., Nichols, B.L., Lin, C., and Hamaker, B.R. 2014. Branch pattern of starch internal structure influences the glucogenesis by mucosal Nt-maltase-glucoamylase. *Carbohydr. Polym.* **111**: 33–40. Elsevier Ltd.
- Naumoff, D.G. 2007. Structure and evolution of the mammalian maltase-glucoamylase and sucrase-isomaltase genes. *Mol. Biol.* **41**(6): 962–973.
- Nichols, B.L., Avery, S., Sen, P., Swallow, D.M., Hahn, D., and Sterchi, E. 2003. The maltase-glucoamylase gene: common ancestry to sucrase-isomaltase with complementary starch digestion activities. *Proc. Natl. Acad. Sci. U. S. A.* **100**(3): 1432–7.
- Nichols, B.L., Eldering, J., Avery, S., Hahn, D., Quaroni, A., and Sterchi, E. 1998. Human Small Intestinal Maltase-glucoamylase cDNA Cloning. *J. Biol. Chem.* **273**(5): 3076–3081.
- Nichols, B.L., Quezada-calvillo, R., Robayo-torres, C.C., Ao, Z., Hamaker, B.R., Butte, N.F., Marini, J., Jahoor, F., and Sterchi, E.E. 2009. Mucosal Maltase-Glucoamylase

Plays a Crucial Role in Starch Digestion and Prandial Glucose Homeostasis of Mice 1 – 3.

- Pontremolil, C., Mozzi, A., Forni, D., Cagliani, R., Pozzoli, U., Menozzi, G., Vertemara, J., Bresolin, N., Clerici, M., and Sironi, M. 2015. Natural Selection at the Brush-Border: Adaptations to Carbohydrate Diets in Humans and Other Mammals. *Genome Biol. Evol.* **7**(9): 2569–2584.
- Price, M.N., Dehal, P.S., and Arkin, a. P. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* **26**(7): 1641–1650.
- Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**(3): e9490.
- Quezada-Calvillo, R., Sim, L., Ao, Z., Hamaker, B.R., Quaroni, A., Brayer, G.D., Sterchi, E.E., Robayo-Torres, C.C., Rose, D.R., and Nichols, B.L. 2008. Luminal starch substrate “brake” on maltase-glucoamylase activity is located within the glucoamylase subunit. *J. Nutr.* **138**(4): 685–92.
- Ren, L., Qin, X., Cao, X., Wang, L., Bai, F., Bai, G., and Shen, Y. 2011. Structural insight into substrate specificity of human intestinal maltase-glucoamylase. *Protein Cell* **2**(10): 827–36.
- Rolfsmeier, M., and Blum, P. 1995. Purification and characterization of a maltase from the extremely thermophilic crenarchaeote *Sulfolobus solfataricus*. These include : Purification and Characterization of a Maltase from the Extremely Thermophilic Crenarchaeote *Sulfolobus solfataricus*. **177**(2): 482–485.
- Sim, L., Quezada-Calvillo, R., Sterchi, E.E., Nichols, B.L., and Rose, D.R. 2008. Human intestinal maltase-glucoamylase: crystal structure of the N-terminal catalytic subunit and basis of inhibition and substrate specificity. *J. Mol. Biol.* **375**(3): 782–92.
- 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**(23): 17763–70.
- Tagami, T., Yamashita, K., Okuyama, M., Mori, H., Yao, M., and Kimura, A. 2013. Molecular basis for the recognition of long-chain substrates by plant alpha-glucosidases. *J. Biol. Chem.* **288**(26): 19296–19303.
- Tan, K., Tesar, C., Wilton, R., Keigher, L., Babnigg, G., and Joachimiak, A. 2010. Novel α -glucosidase from human gut microbiome: substrate specificities and their switch. *FASEB J.* **24**(10): 3939–49.

Draft

Figure 1 Phylogenetic tree of Family GH31 α -glucosidases for experimentally solved structures. Tree is colour coded specifically for eukaryote (red), bacterial (green) and archaea (blue) organisms. Bootstrap values are shown on tree. Substrate preference has been identified for all structures as shown on tree. Preferred linkage(s) derived from starch ($\alpha(1-4)$ and/or $\alpha(1-6)$) are provided with the length of oligosaccharide as identified by the number of glucose residues (G2:maltose, G4:maltotetraose, G6:maltohexaose).

Table 1 Family GH31 α -glucosidases used in the structural analysis of evolutionary changes.

Table 2 Summary of predicted amino acid changes from DIVERGE. Indicated residue number is derived from the original structural sequences.

Figure 2 Predicted evolutionary changes in catalytic region. NT_MGAM (light green), CT_MGAM (grey), NT_SI (red), SBG (violet), MaA (blue), Ro- α G1 (pink) BT_3299 (green) and BT_0339 (golden) are shown in ribbon diagrams. Observed amino acid changes are shown in cyan and the tetrasaccharide substrate acarbose is shown in orange.

Figure 3 Predicted evolutionary changes in looped binding region. NT_MGAM (light green), CT_MGAM (grey), NT_SI (red), SBG (violet), MaA (blue) and Ro- α G1 (pink) are shown in ribbon diagrams. (A). Secondary structural changes predicted to have evolved within Family 31 (B). Observed amino acid changes are shown in cyan and the tetrasaccharide substrate acarbose is shown in orange.

Figure 4 Phylogenetic tree of sequence for characterized Family 31 α -glucosidases. Tree is colour coded specifically for eukaryote (red), bacterial (green) and archaea (blue) organisms. Bootstrap values are shown on tree.

Supplementary Table 1 Characterized Family 31 α -glucosidases used in evolutionary analysis. Table includes organisms the protein belongs to, the common name associated with the protein and the accompanied accession number.

Supplementary Figure 1 Multiple sequence alignment of structurally solved Family 31 α -glucosidases. Signature catalytic sequence WiDMNE is highlighted in red.

Supplementary Table 2 DIVERGE Predictions from Functional Divergence Analysis. Numbering of predictions shown in table come from the reference to the

location within the multiple sequence alignment as shown in Supplementary Figure 1. Highlighted predictions for position 277 and 583 are highlighted in red and are discussed as changes in 'catalytic region' and changes in 'looped binding region' respectively.

Supplementary Figure 2 Multiple sequence alignment of characterized Family 31 α -glucosidases.

Supplementary Figure 3 Predicted eukaryote signature sequence in characterized Family 31 α -glucosidases. (A) Multiple sequence alignment for the region 876-881 with highlighted conserved residues of characterized Family 31 α -glucosidases. (B) NT_MGAM structure (PDB code: 2QMJ) with bound acarbose (orange) and highlighted 'ppY' eukaryote signature sequence (pink).

Draft

Table 1

Organism	PDB ID	Domain	Common Name	Preferred Substrate
<i>Sulfolobus solfataricus</i>	2G3M	Archaea	MalA	Maltose and small maltooligosaccharides (Rolfsmeier and Blum 1995)
<i>Ruminococcus obeum</i>	3MKK	Bacteria	Ro- α G1	Isomaltose (Tan et al. 2010)
<i>Bacteroides thetaiotaomicron</i>	5DJW	Bacteria	BT_3299	Maltose and isomaltose (Chaudet et al. 2012)
<i>Bacteroides thetaiotaomicron</i>	5F7C	Bacteria	BT_0339	Maltose and isomaltose (Chaudet et al. 2012)
<i>Beta vulgaris</i>	3W37	Eukaryota	SBG	Long linear maltooligosacchrides (>G6) (Tagami et al. 2013)
<i>Homo sapiens</i>	2QMJ	Eukaryota	NT_MGAM	Maltose (Sim et al. 2010, Lee et al. 2012)
<i>Homo sapiens</i>	3TOP	Eukaryota	CT_MGAM	Maltose and maltooligosaccharides (Lee et al. 2012)
<i>Homo sapiens</i>	3LPP	Eukaryota	NT_SI	Maltose (Sim et al. 2010, Lee et al. 2012)

Table 2.

Organism	PDB ID	Catalytic Region	Looped Binding Region
<i>Sulfolobus solfataricus</i>	2G3M	Asp346	Ala88
<i>Ruminococcus obeum</i>	3MKK	Asp74	Asp333
<i>Bacteroides thetaiotaomicron</i> -BT_3299	5DJW	Tyr91	Ala349
<i>Bacteroides thetaiotaomicron</i> -BT_0339	5F7C	Gly229	Met474
<i>Beta vulgaris</i>	3W37	Ile233	Asn495
<i>Homo sapiens</i> -NT_MGAM	2QMJ	Thr204	Arg471
<i>Homo sapiens</i> -CT_MGAM	3TOP	Gln1158	His1449
<i>Homo sapiens</i> -NT_SI	3LPP	Gln232	Asp500

Page 23 of 34	Organism	PDB ID	Domain	Common Name	Preferred Substrate
	<i>Sulfolobus solfataricus</i>	2G3M	Archaea	MalA	Maltose and small maltooligosaccharides (Rolfsmeier and Blum 1995)
	<i>Ruminococcus obeum</i>	3MKK	Bacteria	Ro-αG1	Isomaltose (Tan et al. 2010)
	<i>Bacteroides thetaiotaomicron</i>	5DJW	Bacteria	BT_3299	Maltose and isomaltose (Chaudet et al. 2012)
	<i>Bacteroides thetaiotaomicron</i>	5F7C	Bacteria	BT_0339	Maltose and isomaltose (Chaudet et al. 2012)
	<i>Beta vulgaris</i>	3W37	Eukaryota	SBG	Long linear maltooligosacchrides (>G6) (Tagami et al. 2013)
	<i>Homo sapiens</i>	2QMJ	Eukaryota	NT_MGAM	Maltose (Sim et al. 2010, Lee et al. 2012)
	<i>Homo sapiens</i>	3TOP	Eukaryota	CT_MGAM	Maltose and maltooligosaccharides (Lee et al. 2012)
	<i>Homo sapiens</i>	3LPP	Eukaryota	NT_SI	Maltose (Sim et al. 2010, Lee et al. 2012)

Organism	Biochemical ID	PDB ID	Catalytic Region	Looped Binding Region	Flap Region
<i>Sulfolobus solfataricus</i>		2G3M	Asp346		Ala88
<i>Ruminococcus obeum</i>		3MKK	Asp74		Asp333
<i>Bacteroides thetaiotaomicron</i> BT_3299		5DJW	Tyr91		Ala349
<i>Bacteroides thetaiotaomicron</i> BT_0339		5F7C	Gly229		Met474
<i>Beta vulgaris</i>		3W37	Ile233		Asn495
<i>Homo sapiens</i> NT_MGAM	https://mc.manuscriptcentral.com/bcb-pubs	2QMI	Thr204		Arg471
<i>Homo sapiens</i> CT_MGAM		3TOP	Gln1158		His1449
<i>Homo sapiens</i> NT_SI		3LPP	Gln232		Asp500

A

B

PDB ID	Amino Acid
NT-MGAM	Arg471
CT-MGAM	His1449
NT-SI	Asp500
SBG	Asn495
MalA	Ala88
Ro-αG1	Asp333

Common Name	Organism	Accession Number
α -glucosidase 1 (Aglu1;Aglu-1;At5g11720)	<i>Arabidopsis thaliana</i>	AA082630.1
α -glucosidase (α -G I; α -GI)	<i>Aspergillus awamori</i>	BAD06006.1
α -glucosidase B (AgdB)	<i>Aspergillus nidulans</i>	BAB39856.1
α -glucosidase (AgdA;AN2017.2)	<i>Aspergillus nidulans</i>	EAA64849.1
α -glucosidase (Agl;ANG)	<i>Aspergillus niger</i>	BAM72725.1
α -glucosidase (AgdA;Agl; AO090003001209)	<i>Aspergillus oryzae</i>	BAA08125.1
α -glucosidase (AG31A;BspAG31A)	<i>Bacillus sp. SW20</i>	BAQ19546.1
α -glucosidase II / oligo- α -1,6-glucosidase (BthG II)	<i>Bacillus thermoamylolyquefaciens</i>	BAA76396.1
α -glucosidase (BT0339;BT_0339)	<i>Bacteroides thetaiotaomicron</i>	AAO75446.1
α -glucosidase (BT3299;BT_3299)	<i>Bacteroides thetaiotaomicron</i>	AAO78405.1
α -glucosidase (Agl;SBG)	<i>Beta vulgaris</i>	BAM74081.1
α -glucosidase / maltase-glucoamylase (RUMOB_E_03919)	<i>Blautia obeum (Ruminococcus obeum)</i>	ZP_01966167.1
α -glucosidase II α -subunit (BmGII?)	<i>Bombyx mori</i>	BAM78681.1
acidic α -glucosidase	<i>Bos taurus</i>	AAF81637.1
α -glucosidase (Aagr-2:Msp-31;R05F9.12;ZK546.10)	<i>Caenorhabditis elegans</i>	AAA83174.3
α -glucosidase (glucoamylase) (Gca1)	<i>Candida albicans</i>	AAC31968.1
α -glucosidase II (Rot2;CaO19.8589)	<i>Candida albicans SC5314</i>	EAK97887.1
α -glucosidase (Gaa1)	<i>Coturnix japonica</i>	BAA25884.1
α -glucosidase (HdAgl1;HdAgl)	<i>Haliotis discus hannai</i>	BAN67474.1
maltase-glucoamylase (MgaM;Mga;MGAML)	<i>Homo sapiens</i>	AAC39568.2
maltase-glucoamylase (MgaM;Mga;MGAML)	<i>Homo sapiens</i>	AAC39568.2
sucrase-isomaltase (SI)	<i>Homo sapiens</i>	AAT18166.1
sucrase-isomaltase (SI)	<i>Homo sapiens</i>	AAT18166.1
α -glucosidase (Agl1;Agl97)	<i>Hordeum vulgare subsp. vulgare</i>	AAF76254.1
α -glucosidase	<i>Litopenaeus vannamei</i>	CAB85963.1
α -glucosidase (AgdA)	<i>Mortierella alliacea</i>	BAC57563.1
α -glucosidase (lysosomal)	<i>Mus musculus</i>	AAH10210.1
sucrase-isomaltase (SI)	<i>Oryctolagus cuniculus</i>	AAA31459.1
α -glucosidase (Ong1;OngX-H1;Os06g0676700)	<i>Oryza sativa Japonica Group</i>	BAD45913.1
α -glucosidase (Ong2;Ong-2;Ong3;OngX-H3;Os06g067570)	<i>Oryza sativa Japonica Group</i>	BAD45910.1
α -glucosidase (AglA;PTO0092)	<i>Picrophilus torridus</i>	AAT42677.1
α -glucosidase	<i>Pseudozyma tsukubaensis</i>	CAA39501.1
sucrase-isomaltase (SI)	<i>Rattus norvegicus</i>	AAA65097.1
α -glucosidase II (Gls2;Rot2;YBR229c;YBR1526) (ER)	<i>Saccharomyces cerevisiae S288c</i>	CAA85192.1
α -glucosidase (Glz)	<i>Saccharomycopsis fibuligera KZ</i>	CAF31354.1
α -glucosidase (Agl1;Agl;SPAPB24D3.10c)	<i>Schizosaccharomyces pombe 972h-</i>	CAC36906.1
α -glucosidase	<i>Schwanniomyces occidentalis ATCC 2607</i>	BAE20170.1
α -glucosidase (glucoamylase) (Gam1;GAM1)	<i>Schwanniomyces occidentalis ATCC 2607</i>	AAA33923.1
α -glucosidase (MaIA;Saci_1160)	<i>Sulfolobus acidocaldarius DSM 639</i>	AAY80507.1
α -glucosidase (MaIA;SSO3051)	<i>Sulfolobus solfataricus</i>	AAK43151.1
α -glucosidase (ST2525)	<i>Sulfolobus tokodaii</i>	BAB67639.1
α -glucosidase	<i>Thermoanaerobacter ethanolicus JW 20C</i>	ABR26230
α -glucosidase (GluA;TTE0006)	<i>Thermoanaerobacter tengcongensis MB²</i>	AAM23323.1
α -glucosidase (AglA;Ta0298)	<i>Thermoplasma acidophilum</i>	CAC11443.1

<https://mc06.manuscriptcentral.com/bcb-pubs>

Bacteroides thetaiotaomicron-BT_0339
Bombyx mori
Saccharomyces cerevisiae
Candida albicans SC5314
Pseudozyma tsukubaensis
Mucor javanicus
Caenorhabditis elegans
Mortierella alliacea
Schizosaccharomyces pombe
Aspergillus oryzae
Aspergillus niger
Aspergillus awamori
Saccharomycopsis fibuligera
Candida albicans
Schwanniomycetes occidentalis
Sarocladium implicatum
Aspergillus nidulans AgdB
Aspergillus nidulans
Hordeum vulgare subsp. vulgare
Oryza sativa BAD45913
Oryza sativa BAD45910
Beta vulgaris
Arabidopsis thaliana
Litopenaeus vannamei
Homo sapiens CT_MGAM
Rattus norvegicus
Oryctolagus cuniculus
Homo sapiens CT_SI
Homo sapiens NT_MGAM
Homo sapiens NT_SI
Haliotis discus hannai
Coturnix japonica
Mus musculus
Bos taurus
Thermoplasma acidophilum
Ruminococcus obeum
Picrophilus torridus DSM 9790
Sulfolobus acidocaldarius DSM 639
Sulfolobus tokodaii str. 7
Sulfolobus solfataricus
Bacteroides thetaiotaomicron BT_3299
Thermoanaerobacter ethanolicus
Caldanaerobacter subterraneus
Bacillus thermoamyloliquefaciens
Bacillus sp. SW20

-	-	-	-	-	-
N	G	P	-	-	-
D	G	P	-	-	-
N	G	P	-	-	-
N	P	P	Y	A	I
Y	P	K	Y	A	I
M	P	P	Y	K	T
S	I	K	Y	S	I
W	P	P	Y	A	I
H	P	P	Y	V	I
H	P	P	Y	V	I
H	R	P	Y	V	S
Y	P	P	Y	A	I
Y	P	P	Y	A	I
Y	P	P	Y	A	I
Y	P	E	Y	A	I
R	P	P	Y	M	I
S	P	P	Y	S	I
D	P	P	Y	R	I
D	P	P	Y	R	I
D	P	P	Y	R	I
N	P	P	Y	K	I
D	P	P	Y	K	I
S	P	P	Y	P	T
H	P	P	Y	M	P
Y	P	P	V	F	S
Y	P	P	Y	F	P
Y	P	P	Y	F	P
N	P	P	F	T	P
Y	P	P	F	T	P
N	P	P	Y	T	P
S	P	P	Y	T	P
N	P	P	Y	V	P
N	P	P	Y	L	P
-	-	-	-	-	-
A	K	E	F	A	G
-	-	-	-	-	-
-	-	-	-	-	-
-	S	Q	Y	K	-
S	L	P	-	-	-
D	T	P	-	-	-
E	T	P	-	-	-
E	T	P	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

B

Figure 2 Predicted evolutionary changes in catalytic region.
229x233mm (72 x 72 DPI)

Figure 3
276x188mm (72 x 72 DPI)

Biochemistry and Cell Biology