Sequence-structure-function relationships of glycosyltransferases in families GT43, GT47, and GT64

Louis Frederick Lundy Wilson

Jesus College

Department of Biochemistry

Postgraduate School of Life Sciences

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

January 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the word limit prescribed by the Biology Degree Committee.

Declared collaborations:

The IRX9 project (**Chapter 3**) involved little collaboration. However, some of the confocal microscopy was performed by Dr Henry Temple (University of Cambridge).

The EXTL3 project (**Chapter 4**) was collaborative in nature, and involved work performed by Prof. Katrin Mani (Lund University), Dr Steven Hardwick (University of Cambridge), Mr Tom Dendooven (University of Cambridge, now moved to the MRC Laboratory of Molecular Biology, Cambridge), Dr Dima Chirgadze (University of Cambridge), and Dr Theodroa Tryfona (University of Cambridge). Dr Clelton Santos (University of Cambridge, now moved to LNBr/CNPEM) also purified a bacterial enzyme ($BT4658^{GH88}$) specifically for the purposes of this project.

The XAPT/XLPT project (**Chapter 5**) involved little collaboration. However, some transgenic Arabidopsis cell wall material was provided by Dr Li Yu (University of Cambridge).

The wider GT47-A project (**Chapter 6**) involved little collaboration. However, mass spectra were collected by Dr Li Yu, and tomato cell wall material was provided by Dr Yoshihisa Yoshimi (University of Cambridge). Dr Xiaolan Yu (University of Cambridge) and Dr Henry Temple assisted with work to produce the *xlt2 mur3-1* double mutant.

Details of all these collaborations are provided in the main text and materials and methods.

Thesis title: *Sequence-structure-function relationships of glycosyltransferases in families GT43, GT47, and GT64*

Candidate: Louis Frederick Lundy Wilson

Summary

Extracellular carbohydrates are an essential aspect of biology, playing central roles in cell-cell interaction, cell shape, and infectious disease. In eukaryotes, extracellular glycans and glyco-conjugates are synthesised or matured by glycosyltransferase enzymes in the Golgi apparatus. Most glycosyltransferases catalyse the formation of unique glycosidic bonds; thus, the activities of these enzymes are responsible for controlling the diversity of glycan structures. The study of glycosyltransferases can therefore grant insight into the biology of the Golgi apparatus and the ætiology of disease, as well as providing a means to engineer carbohydrate structures for therapeutic or material applications.

GT43 members IRX9 and IRX14 are involved in the synthesis of xylan in Arabidopsis, and are thought to form a multimeric complex with GT47 member IRX10. In this work, I expressed transmembrane fragments of these proteins in *E. coli* and used a reporter assay to demonstrate the sequence-dependent oligomerisation of the IRX9 transmembrane domain. Furthermore, I showed that an IRX9 mutant containing a transmembrane G28I point mutation was unable to complement the phenotype of the *irx9* mutant and appeared to be mislocalised when transiently expressed in tobacco leaves.

In animals, exostosin glycosyltransferases synthesise the backbone of heparan sulphate, and contain both a GT47 and a GT64 domain. Using cryo-EM, I solved the structure of EXTL3, the largest exostosin, in complex with UDP. The structure revealed that the EXTL3 GT47 domain adopts a GT-B fold that has been inactivated by structural changes to the active site. The structure also demonstrated that the GT47 and GT64 active sites are likely separated by a surprising distance in these bi-domain enzymes.

Members of GT47 clade A have particularly diverse activities in plants. I used the structure of EXTL3 to rationalise differences in substrate specificity between these enzymes. Although I was unable to explain the difference in activity between a recently characterised xylan arabinopyranosyltransferase and a homologous xylan galactosyltransferase, I was nevertheless successful in identifying a novel xyloglucan-specific enzyme from *Coffea canephora* on the basis of its protein sequence.

As a whole, the results in this thesis provide insight into the relationship between amino acid sequence, atomic structure, and function in glycosyltransferases. The findings hint at a potential nano-scale organisation of the Golgi apparatus that will be exciting to investigate further.

For my parents

Acknowledgements

First of all, I would like to thank my supervisor Prof. Paul Dupree for his constant support and encouragement over the past five years. I first joined Prof. Dupree's lab as an undergraduate, and ever since I have benefitted from his supreme expertise and diligence in looking after his students. The positive ethos in our lab ultimately derives from Prof. Dupree's enthusiasm for the work of its members. At the same time, I am very grateful for the licence he has given me to steer the course of my own research, throughout which I have also benefitted from the invaluable knowledge and kindness of Prof. Ben Luisi, Prof. Katrin Mani, and Prof. Derek Logan.

I would particularly like to thank Drs Henry Temple, Li Yu, and Jan Łyczakowski for their mentorship, friendship, and support during my PhD. This thesis would not have been possible without them. I also thank Drs Oliver Terrett, Yoshi Yoshimi, Dora Tryfona, Xiaolan Yu, Clelton Santos, Tom Simmons, and Mrs Wendy Gibson for their help and guidance, for which I am extremely grateful. I thank Mr Tom Dendooven, who taught me most of what I know about cryo-EM, as well as Dr Steve Hardwick, Dr Dima Chirgadze, and Mr Lee Cooper for their help with the structural work. I thank Dr Ray Wightman for his knowledge and advice regarding confocal microscopy. I also would like to thank all my fellow students, including Joel, Fede, Katy, Steffi, Keisuke, Rafa, Alberto, Liam, and Konan, for making the lab such a friendly and welcoming environment.

For their help in providing essential materials, I thank Dr Kris Krogh from Novozymes, Drs Sam Brockington, Ángela Cano, Mar Millan, and Mr Alex Summers from the Cambridge University Botanic Garden, Dr Pedro Araújo from the University of Campinas, and Dr Didier Ndeh from Newcastle University. I also thank Prof. Alessandro Senes, Dr Sam Anderson, Prof. George Lomonossoff, and Dr Roger Castells Graells for useful scientific discussions.

I am also very grateful for the financial support I received from the Jesus College Embiricos Hardship Fund, the Department of Plant Sciences, and the University of Cambridge Postgraduate Hardship Lundgren Fund, all of which supported me when it was necessary to extend my PhD due to the COVID-19 pandemic. May continued studies in biochemistry lead to a swift end to this global catastrophe.

I thank my tutors at Jesus College, Prof. David Hanke and Dr Walter Federle, for their mentorship over my time there. Furthermore, I extend my heartfelt thanks to Nessa Skinner,

Alice Kane, and Lottie Barrett-Hague in particular for their counsel and emotional support during my PhD. Finally, I thank my brother George and my parents Elaine and Fred for their patience, encouragement, and general support over the past four years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2-AB	2-aminobenzamide
2-PB	2-picoline-borane
AGP	arabinogalactan protein
AIR	alcohol-insoluble residue
ANTS	8-aminonaphthalene-1,3,6-trisulphonic acid
Ara*f*	L-arabinofuranose
Ara*p*	L-arabinopyranose
BiFC	bimolecular fluorescence complementation
CAZy	Carbohydrate Active Enzymes database (http://www.cazy.org/)
CDG	congenital disorder of glycosylation
CesA	cellulose synthase
CHO	Chinese hamster ovary
CID	collision-induced dissociation
cryo-EM	cryogenic electron microscopy
CS	chondroitin sulphate
CTS	cytoplasmic tail, transmembrane helix, and stem
DHA	3-deoxy-D-*lyxo*-heptulosaric acid
DP	degree of polymerisation
ER	endoplasmic reticulum
EXT	exostosin
FRET	Förster resonance energy transfer
FSC	Fourier shell correlation
Fuc	L-fucose
GAG	glycosaminoglycan
Gal	D-galactose
GalA	D-galacturonic acid
GalNAc	*N*-acetyl-D-galactosamine
GAS	glycine/alanine/serine
GIPC	glycosylinositol phosphorylceramide

Glc	D-glucose
GlcA	D-glucuronic acid
GlcN	D-glucosamine
GlcNAc	*N*-acetyl-D-glucosamine
GPI	glycosylphosphatidylinositol
GT	glycosyltransferase
HEK293	human embryonic kidney 293
HNK	human natural killer
H_nP_m	*n* hexoses and *m* pentoses
HPAEC-PAD	high performance anion-exchange chromatography with pulsed amperometric detection
HS	heparan sulphate
IdoA	L-iduronic acid
IRX	irregular xylem
MALDI-TOF	matrix-assisted laser desorption/ionization-time of flight
Man	D-mannose
MBP	maltose binding protein
MS	mass spectrometry
NDST	*N*-deacetylase/*N*-sulphotransferase
Neu5Ac	*N*-acetylneuraminic acid
NST	nucleotide sugar transporter
OLIMP	oligosaccharide mass profiling
PI	phosphatidylinositol
REO	reducing end oligosaccharide
Rha	L-rhamnose
TMH	transmembrane helix
TU	transcriptional unit
WT	wild type
XAPT	xylan arabinopyranosyltransferase
XLPT	xylan galactopyranosyltransferase
XSC	xylan synthase complex

XyG	xyloglucan
Xyl	D-xylose

Chapter 1 : General introduction

1.1 Glycobiology principles

In addition to a genetic system comprised of nucleic acids, a membrane composed of lipids, and a collection of structural and catalytic proteins, all living cells possess a dense exterior of elaborately structured carbohydrates (Varki, 2011). Moreover, rather than merely decorating the cell, carbohydrates play central roles in molecular recognition, cell organisation, predator evasion, cell structure, and energy storage across the tree of life (Varki, 2017; Solís *et al*, 2015; Schnaar, 2016; Wang *et al*, 2019; Sarkar *et al*, 2009; D'Hulst & Mérida, 2010). Despite this, progress in the study of carbohydrates has been delayed with respect to other biomolecules due to their comparative complexity and lack of decodable template (Roseman, 2001; Varki *et al*, 2009; Lauc *et al*, 2014).

Nevertheless, in recent decades, new techniques have widened our understanding of carbohydrate biology in living organisms (Solís *et al*, 2015), and in addition to the concomitant advancement of basic medical research, the benefits of this are beginning to be felt in more applied fields such as bioenergy, material innovation, and dietary health (Mizrachi *et al*, 2012; Iijima & Hashizume, 2015; Lovegrove *et al*, 2017). Therefore, the continued study of carbohydrates and their biosynthesis has much to offer society. The science regarding the structure, synthesis, biology, evolution, and protein recognition of carbohydrates is often referred to by the name 'glycobiology' (Varki *et al*, 2009).

1.1.1 Describing carbohydrates

Initially, carbohydrates, or 'hydrates of carbon', were defined as molecular species with general formula $C_n(H_2O)_m$. However, as chemists came to understand them better, carbohydrates received a more complex definition: 'polyhydroxyaldehydes', 'polyhydroxyketides', related molecules, and the oligomers or polymers thereof (Robyt, 1998; Sinnott, 2007). Amongst these, the polyhydroxyaldehydes, polyhydroxyketides, and their close derivatives are more commonly known as monosaccharides, and oligomers and polymers of such compounds are therefore termed 'oligosaccharides' and 'polysaccharides' respectively (Berg *et al*, 2002). Furthermore, mono- and oligosaccharides, which include familiar chemical species such as glucose and sucrose, are often also referred to as 'sugars' (Sinnott, 2007). In addition, the term 'glycan' is frequently used to describe oligo- and polysaccharides—particularly in the context

of carbohydrates covalently linked to a non-carbohydrate ('aglycone') to form a 'glycoconjugate' (Varki *et al*, 2009; Moss *et al*, 1995).

1.1.2 Monosaccharide nomenclature

Within monosaccharides, the carbon atoms are linked in a linear unbroken chain. The length of this chain can be used for classification. For instance, trioses contain three carbon atoms, tetroses contain four, pentoses contain five, hexoses contain six, and so on[1] (Sinnott, 2007). That said, monosaccharides may also be classified as aldoses or ketoses according to whether their linear carbon chain contains either an aldehyde or a ketone functional group, respectively (El Khadem, 2012). The carbon atom of such a carbonyl group is given atom number 1 in aldoses, or the lowest possible number in ketoses (following that the carbon assigned number 1 ('C1') must be at one end of the chain, and that the rest of the carbons should be numbered consecutively). Using this system, whichever chiral carbon has the highest atom number is designated the reference carbon, and the arrangement of atoms bonded to it determines the monosaccharide's 'absolute configuration' (Allen & Kisailus, 1992). When displayed as a Fischer projection (Fischer, 1891), oriented with C1 at the top, the monosaccharide's chirality is considered to be dextrorotatory ('D') if the reference carbon's hydroxyl is situated to the right, or lævorotatory ('L') if it is situated to the left (Sinnott, 2007) (**Figure 1.1**).

For monosaccharides of sufficient length, the reactive aldehyde or ketone group permits spontaneous cyclisation; consequently, in solution, monosaccharides exist in an equilibrium between cyclic and acyclic forms (Varki *et al*, 2009). Ring formation transforms the former carbonyl carbon into a new chiral centre—termed the anomeric centre—in one of two possible configurations (El Khadem, 2012). In the cyclic monosaccharide's Fischer projection, if the hydroxyl of this anomeric carbon is situated on the same side as the oxygen bonded to the reference carbon, the configuration is designated 'α', whereas the opposite configuration is designated 'β' (Allen & Kisailus, 1992). An important consequence of this definition is that the two D/L enantiomers of any monosaccharide will also be labelled with opposite anomeric conformations simply by virtue of the nomenclature[2].

[1] These descriptors are not to be confused with similar terms describing oligosaccharide length (for instance the cello-oligosaccharide names cellotriose, cellotetraose, cellopentaose, cellohexaose, &c.).

[2] This is particularly important to remember when comparing β-D-galactose and α-L-arabinopyranose, which are chemically identical save for an additional hydroxymethyl group in the former.

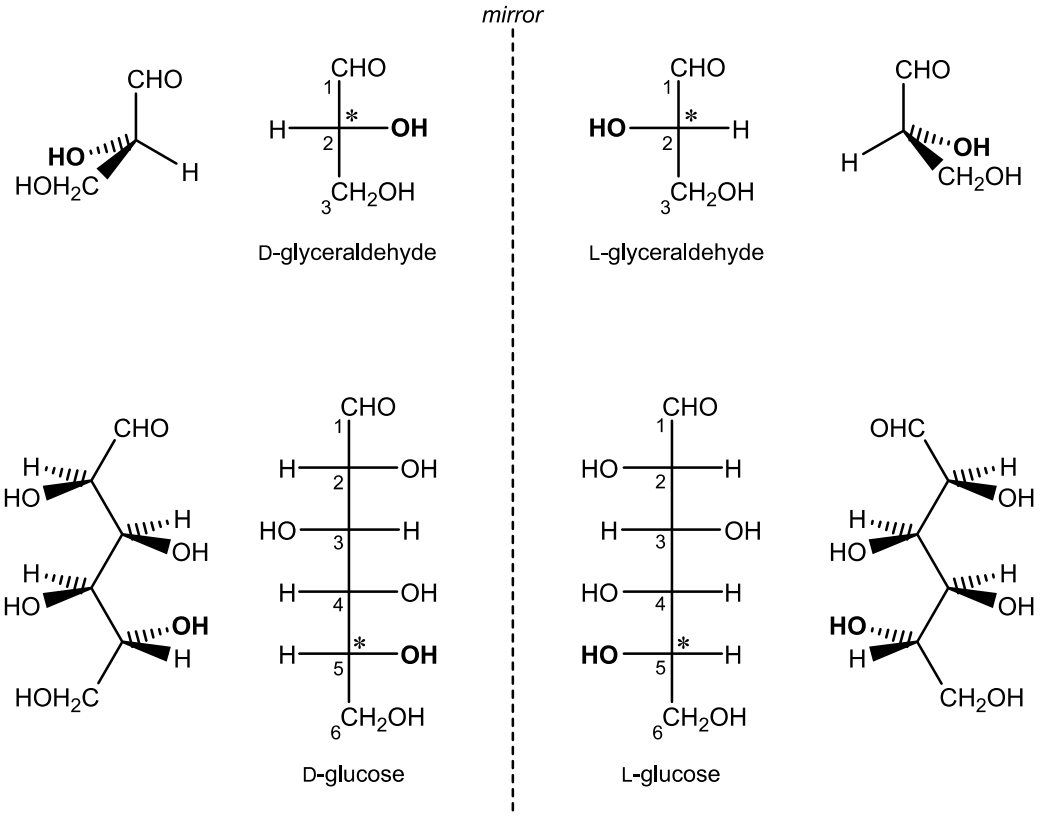


Figure 1.1 'D' and 'L' stereoisomers of monosaccharides can be distinguished by drawing Fischer projections. Fischer projections are drawn such that all horizontal bonds represent bonds pointing towards the viewer. The reference carbon, labelled with '✻', is the highest-numbered asymmetric carbon. If the hydroxyl attached to it appears to its right, the sugar is the D-isomer; if it appears to its left, the sugar is the L-isomer.

Furthermore, since monosaccharides of sufficient size may form either five- and six-membered rings, a third type of isomer can be envisaged (Varki *et al*, 2009). Accordingly, cyclic monosaccharides containing five-membered rings (comprising four carbons and one oxygen) are called 'furanoses', whereas those with six-membered rings (five carbons and one oxygen) are called 'pyranoses' (Sinnott, 2007). While a free aldopentose or aldohexose may therefore adopt a total of five distinct structural conformations in solution (acyclic, α-furanose, β-furanose, α-pyranose, and β-pyranose; **Figure 1.2**), in reality the two pyranose forms predominate in steady state (Robyt, 1998; Sinnott, 2007).

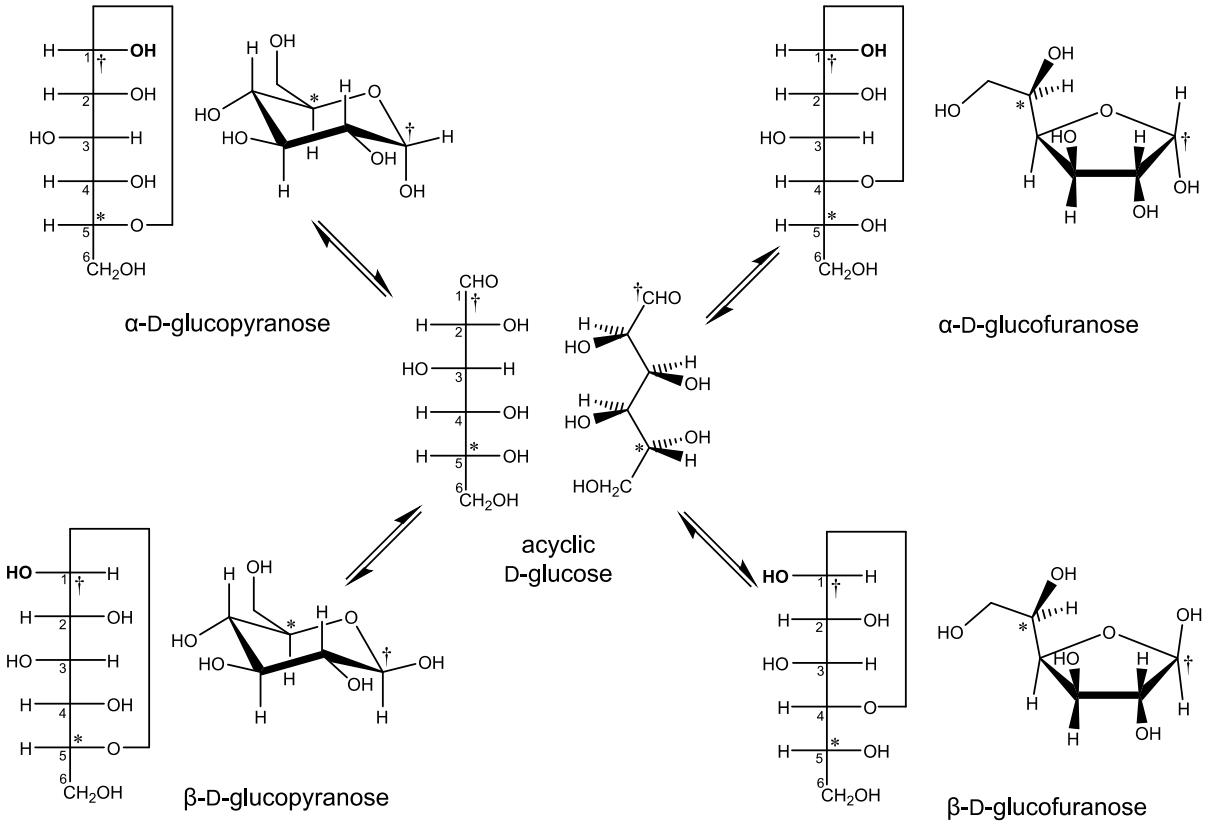


Figure 1.2 Possible conformations created upon the cyclisation of D-glucose. The reference carbon is labelled with '✻', while the anomeric carbon is labelled with '†'. If the hydroxyl of the anomeric carbon is placed on the same side as the oxygen bonded to the reference carbon, the monosaccharide is in the α-anomeric form.

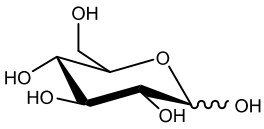
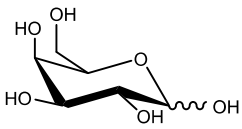
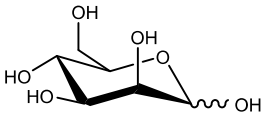
Although further stereoisomers can arise from each of the remaining chiral carbons, inversions at such stereocentres are in fact represented by changing the base name of the monosaccharide, rather than by appending additional symbols or nomenclature. For example, stereochemical inversion at the C2 of β-D-glucopyranose (β-glucose; β-Glc), such that the neighbouring hydrogen and hydroxyl group switch places, results in a monosaccharide that can simply be named β-D-mannopyranose (β-mannose; β-Man) (Varki *et al*, 2009). Stereoisomers such as glucose and mannose are termed 'epimers', and since the epimerisation concerns carbon 2, in this case may specifically be called 'C2 epimers' or '2-epimers' (Sinnott, 2007; Eliasson, 2017).

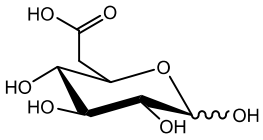
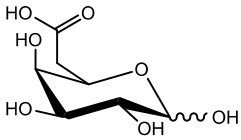
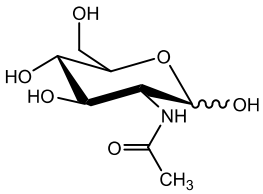
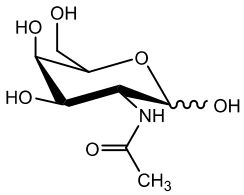
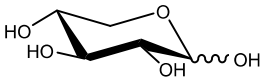
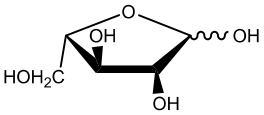
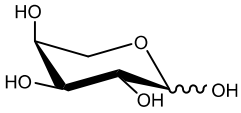
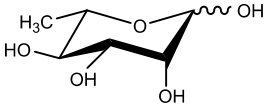
Finally, although many monosaccharides constitute pure polyhydroxyaldehydes or polyhydroxyketones, the definition of 'monosaccharide' often extends to various chemical derivatives of these compounds. For example, replacement of the C6 methoxy group of glucose

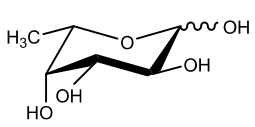
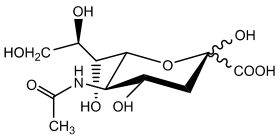
(Glc) with a carboxylic acid group produces a different monosaccharide: glucuronic acid (GlcA) (Sinnott, 2007). Alternatively, if the C2 hydroxyl of glucose is replaced with an amine group, the resulting structure is called glucosamine (GlcN); in turn, acetylation of this amine produces *N*-acetylglucosamine (GlcNAc) (Ma & Gao, 2019). However, when such derivatisation occurs *after* the incorporation of the monosaccharide into oligo- or polysaccharides (for instance in the case of polysaccharide sulphation, methylation, or acetylation), these derivatives are not usually considered to constitute distinct monosaccharides.

In order to standardise representations of glycan structure, each named monosaccharide has been assigned a graphical symbol (Varki *et al*, 2015; Neelamegham *et al*, 2019). For ease of interpretation, the same system will be used in figures throughout this thesis. The subjects of this thesis pertain mainly to the synthesis of glycans found in animals and plants; hence, a list of monosaccharides commonly encountered in animal and plant glycans alongside their structures, symbols, and abbreviations is provided in **Table 1.1**.

Table 1.1 Monosaccharides commonly incorporated into glycans in animals and plants. The anomeric hydroxyl bond is shown as a wavy line to represent both α and β conformations. Sugar-nucleotides used in glycan synthesis are listed; *cytosol only, †plastid only.

Trivial name	Specific configuration	Abbrev.	Structure	Symbol	Sugar-nucleotide
glucose	D-glucopyranose	Glc			UDP-Glc GDP-Glc* ADP-Glc†
galactose	D-galactopyranose	Gal			UDP-Gal
mannose	D-mannopyranose	Man			GDP-Man

glucuronic acid	D-glucuronopyranose	GlcA			UDP-GlcA
galacturonic acid	D-galacturonopyranose	GalA			UDP-GalA
N-acetyl-glucosamine	*N*-acetyl-D-glucosaminopyranose	GlcNAc			UDP-GlcNAc
N-acetyl-galactosamine	*N*-acetyl-D-galactosaminopyranose	GalNAc			UDP-GalNAc
xylose	D-xylopyranose	Xyl			UDP-Xyl
arabinofuranose	L-arabinofuranose	Ara*f*			UDP-Ara*f*
arabinopyranose	L-arabinopyranose	Ara*p*			UDP-Ara*p*
rhamnose	L-rhamnopyranose	Rha			UDP-Rha

fucose	L-fucopyranose	Fuc			GDP-Fuc
N-acetyl-neuraminic acid		Neu5Ac			CMP-Neu5Ac

1.1.3 Glycosidic bonds

In cells, a large proportion of monosaccharides are incorporated into oligosaccharides, polysaccharides, and glycoconjugates thereof. Within these, the monosaccharides are linked to each other by glycosidic bonds, formed by the condensation reaction between the anomeric hydroxyl of one monosaccharide (i.e. the C1 hydroxyl in an aldose) and any of the particular hydroxyls of another (Berg *et al*, 2002; Rao *et al*, 1998). When a monosaccharide's anomeric oxygen participates in such a bond, the monosaccharide becomes fixed in an 'α' or 'β' cyclic conformation, thereby losing the potential reducing power of its aldehyde or ketone group. Hence, the termini of oligo- and polysaccharides can be differentiated using the labels 'reducing end' (constituting the only monosaccharide unit with an unsubstituted anomeric hydroxyl) and 'non-reducing end' (Varki *et al*, 2009).

These glycosidic bonds, or 'linkages', can therefore be found in a range of flavours when accounting for the anomeric conformation of the first monosaccharide and the choice of hydroxyl in the second. When combined also with the number of different monosaccharide building blocks, this permits enormous variety in carbohydrate structure—it has been estimated that a trimer of three different hexoses can be constructed in somewhere between 1,056 and 27,648 different ways alone, for instance (Varki *et al*, 2009). Especially in light of the possibility of glycan branching (also called 'substitution'), carbohydrate polymers can therefore comprise many more unique structures than polypeptides or polynucleotides of equivalent length (Solís *et al*, 2015). This unparalleled level of biomolecular information underlies the dynamics of many cell-cell interactions, including those of pathogen-host interactions (Varki, 2011; Kreisman & Cobb, 2012).

Furthermore, even a single change to the type of linkage can majorly affect the properties of a polysaccharide: for example, although amylose and cellulose are both homopolymers of glucose, the former, which contains α1,4 linkages, forms helices and random coils in solution (Pérez & Bertoft, 2010), whereas the latter, which contains only β1,4 linkages, assembles into crystalline microfibrils that can contain as many as 80 chains in cotton—or that can reach a thickness of 15 nm in certain species of tunicate (Jarvis, 2018) (see **Figure 1.3** for comparison of α and β glycosidic bonds). Such properties are central to the functions of these polysaccharides in energy storage and cell structure, respectively.

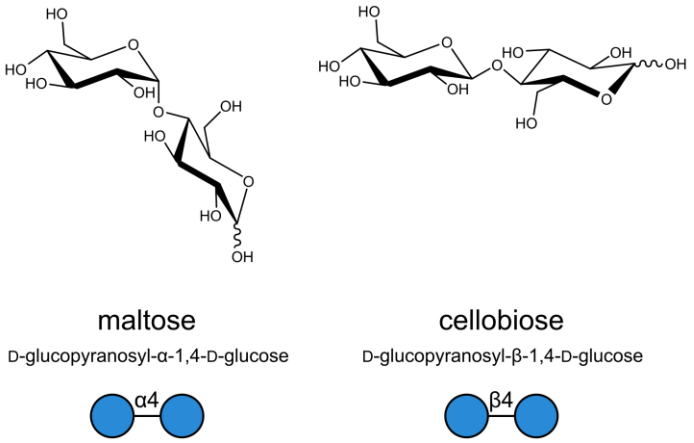


Figure 1.3 The anomeric configuration of a glycosidic bond can impart major differences in overall conformation for oligo- and polysaccharides.

1.2 Eukaryotic glycans

In eukaryotes, glycans and glyco-conjugates comprise a wide range of structures, including glycoproteins, glycolipids, and polysaccharides. Such molecules are largely to be found at the cell surface, in the extracellular space, and in the lumen of secretory pathway organelles. Though glycosylation is often initiated in the endoplasmic reticulum (ER), and some simple structures are produced by cytoplasmic activities, the overwhelming majority of glycans are constructed and matured in the Golgi apparatus (Stanley, 2011; Litwack, 2018).

1.2.1 Glycoproteins

The most widely conserved type of glycan found in eukaryotes comprises those in *N*-glycosylated proteins, which are also present in Archæa and a limited number of bacteria (Varki *et al*, 2009; Jarrell *et al*, 2014). In virtually all eukaryotes, *N*-glycosylation begins with the transfer of a $Glc_3Man_9GlcNAc_2$ oligosaccharide to a protein acceptor (specifically the asparagine sidechain of a universal N-X-S/T motif) in the ER, and the subsequent trimming and/or extension of this glycan is critical for progression of the glycoprotein through the secretory pathway (Aebi, 2013).

Though less well conserved in structure, *O*-glycosylation of hydroxyl-containing sidechains is also extremely common in eukaryotes. In fungi, *O*-glycans are frequently attached through mannosyl residues to serine and threonine sidechains in secretory proteins for stabilisation purposes (Goto, 2007), whereas in plants, many important cell wall proteins (such as arabinogalactan proteins (AGPs) and extensins) are *O*-glycosylated with complex carbohydrate moieties through hydroxyproline arabinosylation or galactosylation (Held *et al*, 2015). However, the most widespread form of *O*-glycosylation in eukaryotes is mucin-type *O*-glycosylation, which is initiated by the transfer of GalNAc to a serine or threonine residue (Hang & Bertozzi, 2005). Mucins are abundant in animals, and, amongst other functions, serve to protect epithelial surfaces from infection and physical damage (Pelaseyed *et al*, 2014; Corfield, 2015). The ABO and Lewis blood group antigens in humans are also mucin-type *O*-glycans (Schnaar, 2016).

1.2.2 Extracellular polysaccharides

In animals, proteoglycans make up another major class of glycoconjugate. These molecules consist of a polypeptide core decorated with one or more (often *O*-linked) glycosaminoglycan chains (Couchman & Pataki, 2012; Kolset *et al*, 2004). Glycosaminoglycans (GAGs) are long, usually unbranched polysaccharides that consist of alternating hexosaminyl and hexosyl/hexuronosyl residues (Varki *et al*, 2009; Pomin, 2014). Glycosaminoglycans can be sorted into four classes: heparan sulphate and heparin, which contain alternating GlcNAc and GlcA or iduronosyl (IdoA) residues; chondroitin sulphate and dermatan sulphate, which contain alternating GlcNAc and GalA or IdoA (respectively); keratan sulphate, which contains alternating GlcNAc and Gal; and hyaluronan, which contains alternating GlcNAc and GlcA (albeit with a different linkage to heparan) (Zhang *et al*, 2010). Unlike the other glycosaminoglycans, which are synthesised in the Golgi, hyaluronan is made at the plasma membrane and is not linked to a protein core (Tammi *et al*, 2011). A large number of specific glycosaminoglycan-binding proteins can be found in the extracellular matrix, and glycosaminoglycans are highly essential for many processes involving cell-cell interactions such as growth and development (Varki *et al*, 2009; Couchman & Pataki, 2012). Finally, though not itself a glycosaminoglycan, one further *O*-linked polysaccharide is known to be important in animal cell surface interactions: matriglycan. This polysaccharide contains an alternating Xyl and GlcA repeat, and helps to link dystroglycan to laminin in the extracellular matrix; its abrogation is one cause of Duchenne muscular dystrophy (Yoshida-Moriguchi & Campbell, 2014).

Plants are not known to possess glycosaminoglycans. Nevertheless, the majority of the plant cell wall is made up of a variable collection of neutral and anionic polysaccharides (Scheller & Ulvskov, 2010). Cellulose, in particular, can be found in all plant and algal lineages; it is thought that an ancient alga acquired the machinery for its synthesis by gene transfer from an endosymbiotic cyanobacterium (Popper *et al*, 2011). The β1,4-linked glucan chains of cellulose are assembled into rigid microfibrils whose deposition underlies both the strength and the extensibility of the cell wall (Thomas *et al*, 2013).

The prevalence of other polysaccharides in plant cell walls depends on species and wall type. Two major types of cell wall are recognised in the model plant, Arabidopsis: primary walls, which are made by all cells during growth, and secondary walls, which are a more rigid type of wall produced to reinforce certain tissues after growth has ceased (Cosgrove & Jarvis, 2012). In primary walls and the middle lamella, 'pectic' polysaccharides rich in α-GalA are prevalent (Atmodjo *et al*, 2013; Held *et al*, 2015). This 'pectin' forms semi-rigid gels held together both by ionic interactions mediated by calcium ions (which can be modulated through methylesterification of backbone galacturonosyl residues) and by covalent borate diester crosslinks (Fry, 1986, 2011; Gawkowska *et al*, 2018). Other polysaccharides commonly found in primary walls include xyloglucan (which possesses a β1,4-linked glucan backbone decorated with xylose and various secondary substitutions), (galacto)glucomannan (consisting of a backbone of β1,4-linked mannosyl and glucosyl residues that may be substituted with galactose), and mixed-linkage glucan (an unbranched glucan with both β1,4-linkages and β1,3-linkages). In secondary cell walls and the cell walls of grasses, xylan (a β1,4-linked polymer of xylosyl residues often substituted with glucuronosyl and/or arabinosyl residues) is the main non-cellulosic polysaccharide (Scheller & Ulvskov, 2010; Rennie & Scheller, 2014). While the precise role of many of these polysaccharides is yet to be determined, several of them, particularly xylan, are thought to bind directly to cellulose microfibrils (Simmons *et al*, 2016; Cosgrove, 2014; Yu *et al*, 2018).

The fungal cell wall differs significantly from the plant cell wall in its polysaccharide composition. The main cell wall polysaccharide in fungi is β1,3-glucan; chitin (made up of β1,4-linked GlcNAc residues) and β1,6-glucan are also prominent (Varki *et al*, 2009). In particular, chitin—which can also be found in many invertebrate animals—assembles into microfibrils in a fashion reminiscent of cellulose (Merzendorfer, 2011).

1.2.3 Storage glycans

Although cellulose is the most abundant organic compound on the planet (Kamide, 2005), probably the most familiar polymeric carbohydrate encountered in daily life is that of vegetable starch. Synthesised exclusively in plastid organelles, starch is the main storage carbohydrate in vascular plants, and is made up primarily of two homopolymers of glucose: amylose and amylopectin (Bahaji *et al*, 2014). Amylose contains almost exclusively α1,4 linkages, and can thus form stable helices; amylopectin, on the other hand, contains linkages that are roughly 95% α1,4 and 5% α1,6, and therefore prefers to form gels and films (Pérez & Bertoft, 2010). Amylose has also been detected in some fungi (Gorin & Barreto-Bergter, 1983).

Nonetheless, in animals, fungi, and in fact many bacteria, the main storage polysaccharide is glycogen (Takahara & Matsuda, 1976). Glycogen is similar in composition to amylopectin, but is more highly branched—with one α1,6 branch for every ten glucosyl residues in animals (Calder, 1991). Glycogen is synthesised in the cytoplasm and is attached to a core protein, glycogenin, through an unusual Glc-*O*-Tyr linkage (Litwack, 2018).

1.2.4 Glycolipids

Lipids have previously been sorted into eight classes: fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, saccharolipids, polyketides, sterol lipids, and prenol lipids (Fahy *et al*, 2005). Lipids in all major classes can be found in glycosylated forms (Yang & Tang, 2000). Being virtually ubiquitous in eukaryotes (Kinoshita *et al*, 1997), glycosylphosphatidylinositol (GPI) is the best conserved glycolipid amongst these. The glycan present in this glycerophospholipid possesses a highly conserved Man_3GlcN-*myo*-inositol core that is linked via phosphoethanolamine to a peripheral membrane protein—this connection allows GPIs to carry out their function as membrane anchors (Paulick & Bertozzi, 2008). Within GPIs, the lipid-glycan connection itself is usually between *myo*-inositol and the phosphate group of phosphatidylinositol (PI) (Maeda *et al*, 2010); however, unphosphorylated glycerolipids can also be directly glycosylated—in animal spermatogenic cell seminolipids for instance (Honke, 2013).

In plants and fungi, this same conserved glycan can be transferred to phosphoceramide, a sphingolipid (Maeda *et al*, 2010). In fact, glycosylated sphingolipids tend to make up a significant proportion of glycolipids in eukaryotes. In plants and fungi, glycosylinositol phosphorylceramides (GIPCs) are common, perhaps making up 25% of total lipids in the plant plasma membrane (Gronnier *et al*, 2016; Marinas *et al*, 2010; Fang *et al*, 2016). GIPCs appear

to play a role in the formation of microdomains in the plasma membrane (Borner *et al*, 2005; Cacas *et al*, 2016).

Directly glycosylated ceramides such as glucosylceramide, on the other hand, are found across the eukaryotic lineage, and are the major form of glycolipid in animals (Leipelt *et al*, 2001; Varki *et al*, 2009). Gangliosides, for instance, are highly expressed in animal brain. These glycolipids can have particularly complex glycans, and possess a large variety of functions, often by virtue of specific interactions with molecules in the extracellular space and in the membrane itself (Kolter, 2012).

Many of the glycans summarised in this section are illustrated in **Figure 1.4**.

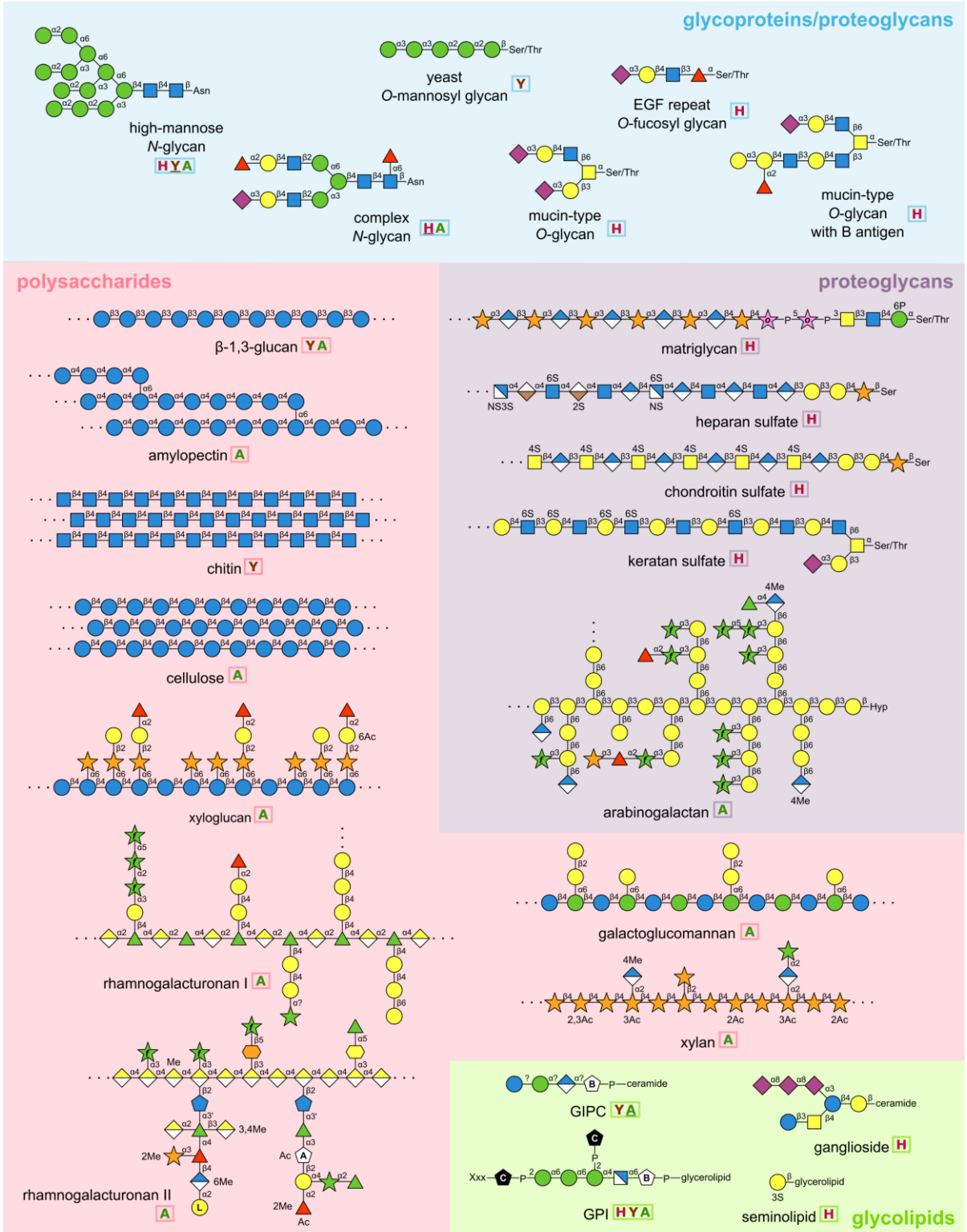


Figure 1.4 Examples of glycan structures in eukaryotes. H: present in *Homo sapiens*, **Y**: present in *Saccharomyces cerevisiæ*, **A**: present in *Arabidopsis thaliana*.

1.3 Glycosyltransferases

It has been traditionally thought that, for every type of glycosidic linkage present in an organism, an individual enzyme activity is required to produce it (Roseman, 2001; Varki *et al*, 2009). Such enzymes are called glycosyltransferases, and accordingly they make up 1–2% of a typical genome in any branch of life (Lairson *et al*, 2008). Glycosyltransferases (GTs) catalyse the transfer of a monosaccharide, or sometimes an oligosaccharide, from an 'activated' (i.e. exergonically hydrolysable) sugar donor to an acceptor, which may be carbohydrate or aglycone in nature (Breton *et al*, 2006; Stetten, 1960). While sugar donors can be phosphosugars or lipid-phosphosugars, most commonly they are nucleotide sugars, in which the sugar is usually α-linked at its C1 hydroxyl to the β-phosphate of a nucleoside diphosphate (Lairson *et al*, 2008; Varki *et al*, 2009) (**Figure 1.5a**). Due to the nature of nucleotide sugar synthesis pathways, the particular nucleotide used varies depending on the monosaccharide (Bülter & Elling, 1999) (see **Table 1.1** for a list of common nucleotide sugars).

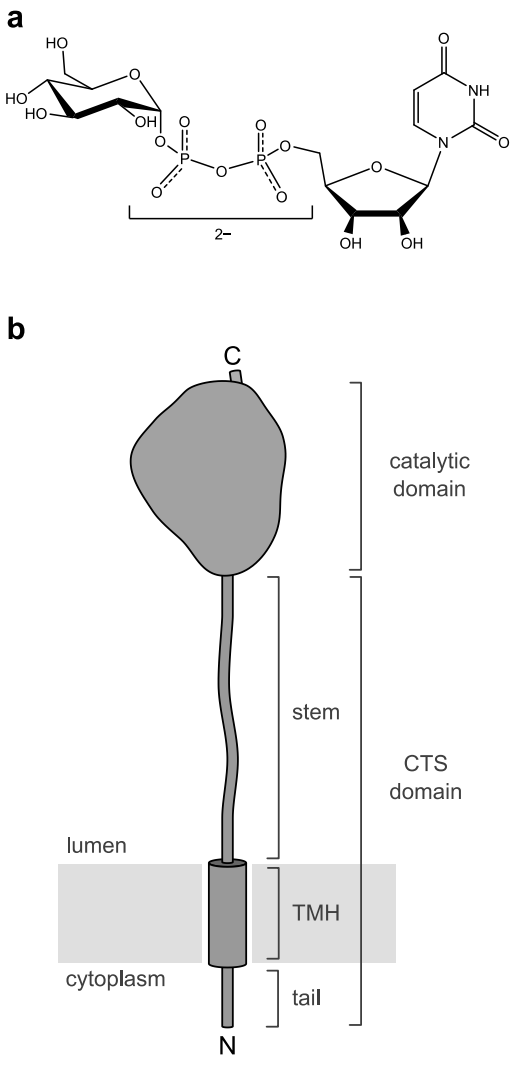


Figure 1.5 Most glycosyltransferases use nucleotide sugar donors. a Structure of UDP-glucose, a nucleotide sugar. **b** Domain structure of a typical Golgi GT.

In eukaryotes, the majority of GTs are found in the Golgi (Stanley, 2011). Hence, to avoid secretion from the cell, GTs must be anchored to the Golgi membrane. Accordingly, the most common architecture of a Golgi GT is that of a Type II single-pass transmembrane protein, with a short N-terminal cytoplasmic tail, a single transmembrane helix, a disordered stem region, and a globular C-terminal catalytic domain in the Golgi lumen (Breton *et al*, 2006; Welch & Munro, 2019) (**Figure 1.5b**); the cytoplasmic tail, transmembrane helix, and stem are often collectively referred to as the 'CTS' domain (Tu & Banfield, 2010). Golgi GTs frequently form oligomers—most commonly homodimers—that can often be disulphide-linked, and

interactions can be mediated by different parts of the protein depending on the GT (Young, 2004; Kellokumpu *et al*, 2016; Harrus *et al*, 2018).

1.3.1 Catalytic domains of glycosyltransferases

Naturally, the active site of a glycosyltransferase can be found within its catalytic domain. This domain, sometimes called the glycosyltransferase domain (GT domain), is also responsible for binding the donor sugar and the acceptor (Kapitonov & Yu, 1999); other domains are not normally involved in catalysis. GTs that use nucleotide sugars as donors are known as Leloir glycosyltransferases (Lairson *et al*, 2008).

Depending on whether their catalytic mechanism results in the inversion of the transferred sugar's anomeric centre, GTs can be divided into two classes: inverting and retaining (Breton *et al*, 2006). GT domains have also been sorted into a wide range of families based on sequence similarity (Campbell *et al*, 1997; Coutinho *et al*, 2003), and at the beginning of 2021, at least 111 GT families have been listed on the Carbohydrate Active Enzymes database (CAZy) server (http://www.cazy.org/; Lombard *et al*, 2013). Generally, GTs from the same family share the same stereoselectivity, though there are a number of families that contain both inverting and retaining GTs (Lairson *et al*, 2008). Despite this, GT activity is difficult to predict from sequence due to the subtle nature of GT substrate specificity (Breton *et al*, 2006; Gloster, 2014), though recent progress has been made in this field using machine learning techniques (Yang *et al*, 2018; Taujale *et al*, 2020). Furthermore, although donor specificities have been successfully modified through point mutation, some of these mutations have actually been distal to the active site and there seem to be no general rules for engineering substrate specificity (Lairson *et al*, 2008; Chang *et al*, 2011).

Although GTs from different CAZy families share extremely little sequence similarity, structural data have revealed that GTs adopt only a small number of folds (Breton *et al*, 2006). Almost all characterised GTs adopt a GT-A fold, a GT-B fold, or a multi-span GT-C fold, though several 'orphaned' GTs that do not fit into these categories have also been described (Lairson *et al*, 2008). In recent years, founding members of (exclusively bacterial) GT-D and GT-E folds have also been proposed (Zhang *et al*, 2014; Kattke *et al*, 2019).

Of the folds observed in GT structures to date, the GT-A fold is the most common, and is found in all kingdoms of life (Lairson *et al*, 2008; Taujale *et al*, 2020). The alternating α-helices and β-strands of the GT-A secondary structure fold into a three-layered sandwich in which the central β-sheet is surrounded by α-helices on both sides (an α/β/α sandwich; **Figure 1.6**)

(Breton *et al*, 2006). This central sheet typically contains seven strands and possesses a 3214657 topology, with strand 6 lying antiparallel to the otherwise parallel configuration; in addition, a small flanking antiparallel sheet is formed from two further strands (4' and 7') (Breton *et al*, 2006; Chang *et al*, 2011). The overall secondary structure is similar to the Rossmann fold, a common structural motif with 321456 sheet topology, although the GT-A fold has in fact been described as two closely abutting Rossmann-like folds (corresponding to 3214 and 657) (Breton *et al*, 2006; Lairson *et al*, 2008). Most GT-A glycosyltransferase activities are metal cation-dependent; thus, a metal-coordinating DxD motif is usually found in the loop connecting β4 to β4' at the junction between the two subdomains (Wiggins & Munro, 1998; Breton *et al*, 2006; Taujale *et al*, 2020). This metal ion, typically manganese or magnesium, helps to bind the negatively charged phosphates of nucleotide sugars, and in metal-independent GT-A enzymes is often substituted with cationic side chains (Lairson *et al*, 2008). Besides this interaction, the nucleotide sugar is bound primarily by the N-terminal half of the GT-A fold, and often residues at the end of sheet β1 are involved in nucleobase recognition (Breton *et al*, 2006; Chang *et al*, 2011). The C-terminal subdomain is much more variable, in line with its greater role in acceptor binding, though a central region comprising β6 and the two α-helices that follow it is usually conserved (Breton *et al*, 2006; Chang *et al*, 2011). Nevertheless, in metal-dependent GT-As, a histidine close to the C-terminus (the 'C-His') is often present, and helps to co-ordinate the metal ion (Taujale *et al*, 2020). However, in general, binding of both donor and acceptor is achieved mainly through variable loops that are inserted into the more conserved fold (Moremen & Haltiwanger, 2019; Taujale *et al*, 2020). Often, conformational changes induced by nucleotide sugar binding are necessary for the subsequent binding of the acceptor (Breton *et al*, 2006).

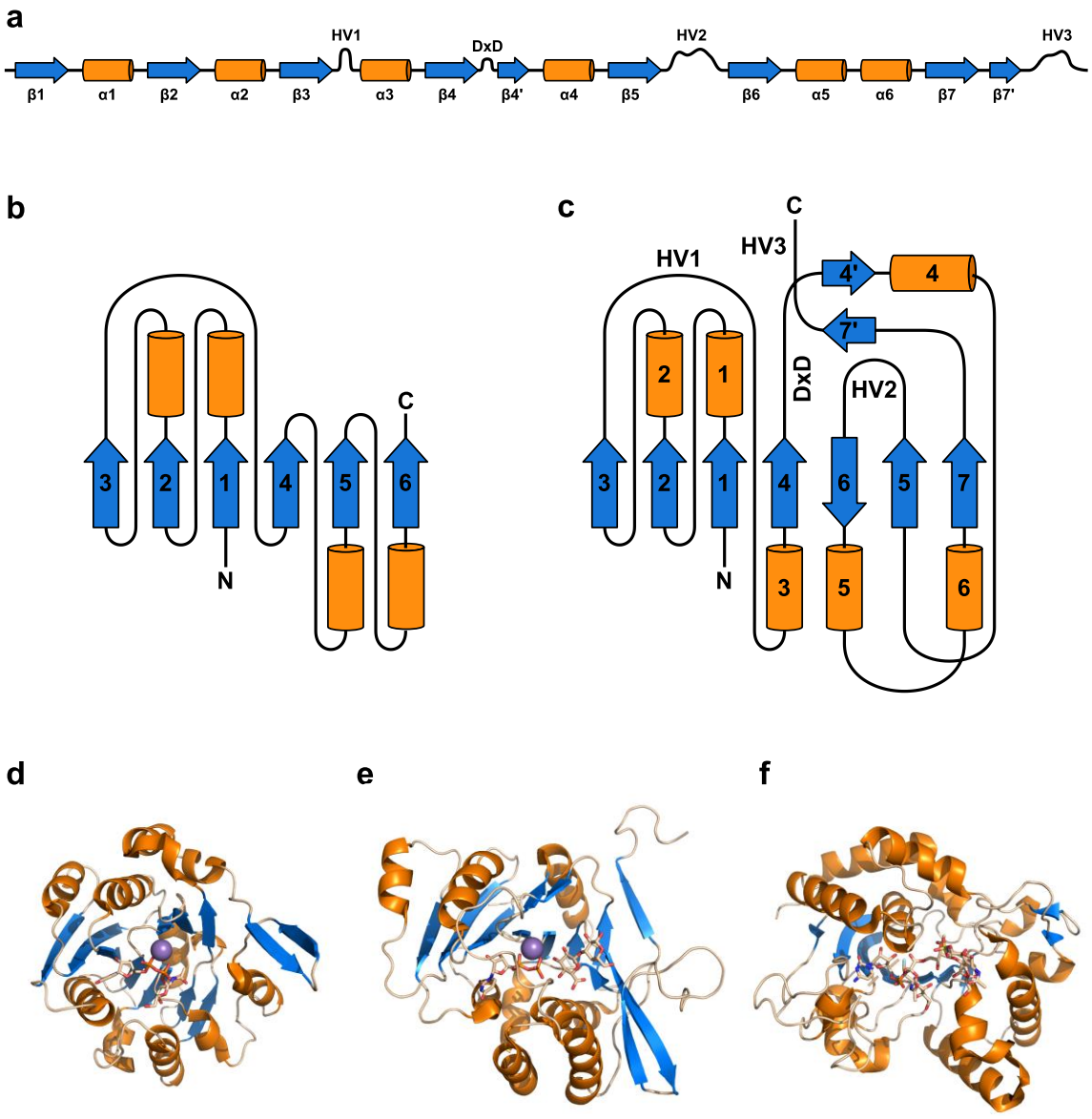


Figure 1.6 Secondary and tertiary structure of the GT-A fold. a Archetypal secondary structure of a GT-A glycosyltransferase (adapted from Taujale *et al*, 2020, with numbering as per Breton *et al*, 2006 and Chang *et al*, 2011). HV: hypervariable region. **b** Topology of the Rossmann fold. **c** Typical topology of a GT-A fold glycosyltransferase (broadly based on Chang *et al*, 2011). **d** Crystal structure (PDB: 2Z87) of the N-terminal domain of *E. coli* K4 chondroitin synthase, an (inverting) GT2 family member, in complex with UDP-GalNAc and Mn^{2+}. **e** Crystal structure (PDB: 1ON8) of mouse EXTL2, a (retaining) GT64 member, in complex with UDP, Mn^{2+}, and a disaccharide acceptor. **f** Crystal structure (PDB: 5BO9) of the human sialyltransferase ST8SiaIII (GT29; inverting) in complex with CMP-3FNeu5Ac and a trisaccharide acceptor. Although this GT29 member adopts a GT-A fold, it has lost many sequence features typical of GT-A-fold enzymes.

GT2 comprises the largest GT-A-fold CAZy family, and has been proposed to represent the evolutionary progenitor of the other GT-A families (Lairson *et al*, 2008). Indeed, a recent global analysis of GT-A structures suggested that the N-terminal GT2 domain of the *E. coli* K4 chondroitin synthase (PDB: 2Z87) is the closest structure yet solved to the GT-A consensus (Taujale *et al*, 2020). In the same work, it was proposed that inverting and retaining activities have evolved independently several times within the GT-A superfamily. Nevertheless, the actual mechanism of catalysis appears to be conserved within each type. Inverting GT-A glycosyltransferases, which include the GT2 family, are thought to use an S_N2-type mechanism, and generally use a Asp, Glu, or His catalytic base found in an xED motif located at the N-terminus of the second α-helix of the C-terminal conserved region (α6) (Gloster, 2014; Moremen & Haltiwanger, 2019; Taujale *et al*, 2020). The mechanism of retaining GT-As, however, has proved to be a controversial topic in the past; nonetheless, recent evidence seems to support an 'S_Ni-like' mechanism involving the spontaneous lysis of the nucleotide—sugar linkage to form a short-lived oxocarbenium intermediate that can be attacked by the acceptor from the same face as the leaving nucleotide (Ardèvol *et al*, 2016; Albesa-Jové *et al*, 2019; Moremen & Haltiwanger, 2019). It has been suggested that the evolutionary transition from inverting to retaining activity is brought about by the loss or repositioning of the catalytic base necessary for the former (Taujale *et al*, 2020).

Many other glycosyltransferase domains adopt the GT-B fold, which, like GT-A, can also be found in all kingdoms of life (Lairson *et al*, 2008). The GT-B superfamily contains the large GT1 family, which has attracted much interest due to the role of its members in natural product synthesis (Bock, 2016). However, it is thought that GT-Bs arose from a yet larger family: GT4 (Lairson *et al*, 2008; Breton *et al*, 2012). As with GT-A, the secondary structures of GT-B glycosyltransferases consist mainly of alternating α-helices and β-strands. However, in contrast to GT-A, the GT-B fold consists of two loosely connected complete Rossmann-fold domains with a C-terminal linker region and an active site situated in the cleft between the two domains (**Figure 1.7**) (Breton *et al*, 2006; Albesa-Jové *et al*, 2014; Gloster, 2014). The central sheet of each domain is typically six- or seven-stranded, and normally possesses 321456(7) topology (Liu & Mushegian, 2003; Chang *et al*, 2011). The C-terminal domain is responsible for binding the donor, and is reasonably well conserved between different GT families, whereas the N-terminal domain makes most of the contacts with the acceptor and varies substantially from family to family (Breton *et al*, 2006; Chang *et al*, 2011; Albesa-Jové *et al*, 2014). In some cases, substantial flexibility between the two domains has been demonstrated by

crystallography, and it has been suggested that the deep cleft of the GT-B fold is more well suited than the GT-A fold to binding bulky acceptors such as heavily branched glycans and globular proteins (Albesa-Jové *et al*, 2014; Moremen & Haltiwanger, 2019).

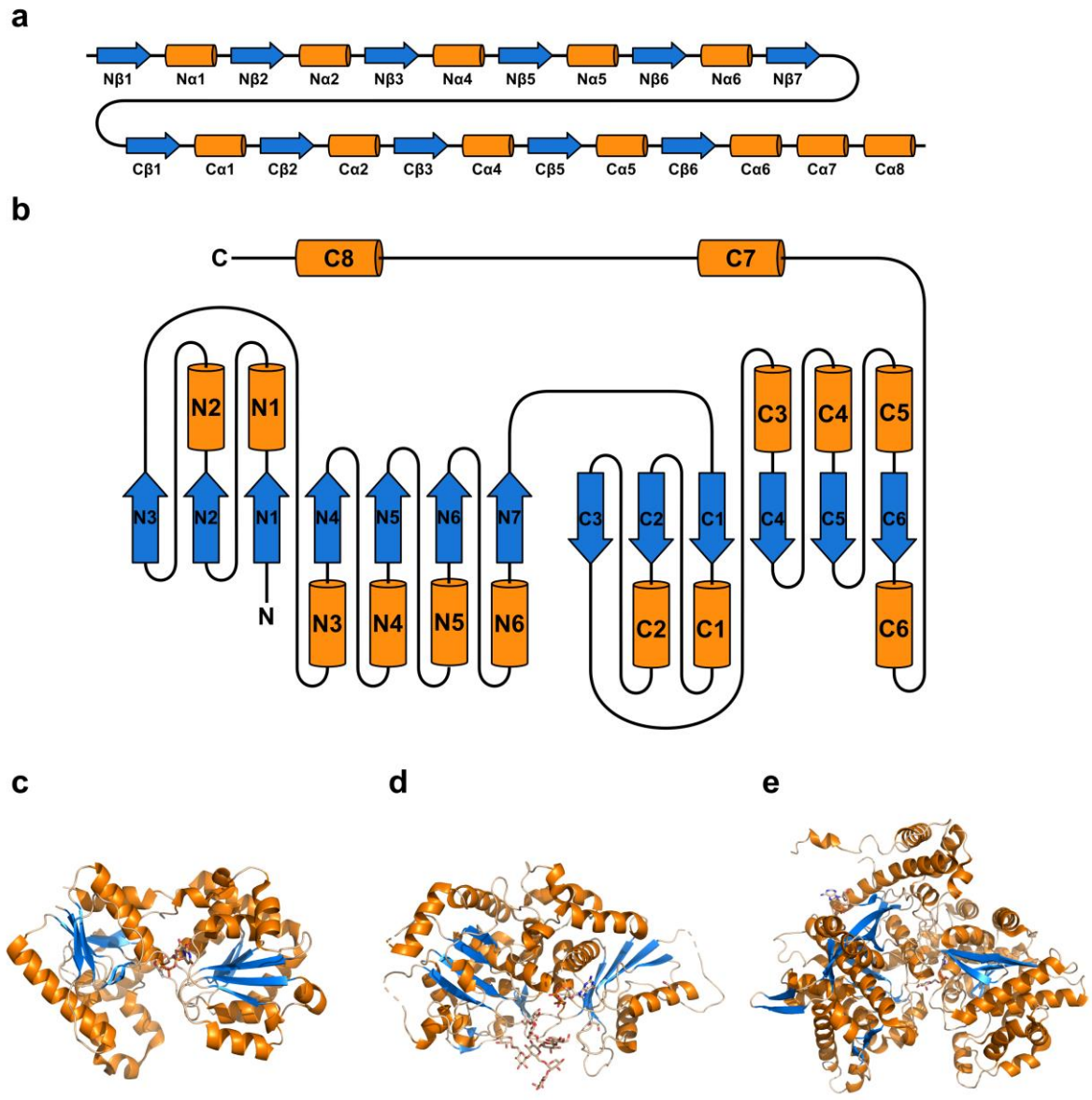


Figure 1.7 Secondary and tertiary structure of the GT-B fold. a Archetypal secondary structure of a GT-B glycosyltransferase (using numbering from Chang *et al*, 2011). **b** Typical topology of a GT-B glycosyltransferase (based on Chang *et al*, 2011). **c** Crystal structure (PDB: 2IW1) of *E. coli* WaaG (GT4; retaining) in complex with UDP-2FGlc. **d** Crystal structure (PDB: 5KOR) of Arabidopsis FUT1 (GT37; inverting) in complex with UDP and a nonasaccharide acceptor. **e** Crystal structure (PDB: 1FA9) of glycogen phosphorylase from human liver. Though not a glycosyltransferase, this enzyme is related to GT-Bs and is assigned to the GT35 family.

GT-B enzymes are metal ion-independent. As in other Rossmann-fold proteins, these enzymes rely on the α-helix dipole effect at the *N*-terminus of helix Cα4 to stabilise phosphate binding—though a few GT-B activities are accelerated in the presence of certain cations nonetheless (Hol *et al*, 1978; Hu & Walker, 2002). Accordingly, there appear to be no universally conserved amino acid motifs in GT-B enzymes (Hu & Walker, 2002). That being said, a glycogen phosphorylase-like motif has been found in the C-terminal domains of many GT-Bs, involving a conserved Asp/Glu in helix Cα4 that typically forms a hydrogen bond to the ribose moiety of the nucleotide sugar (Wrabl & Grishin, 2001; Hu & Walker, 2002; Breton *et al*, 2006; Martinez-Fleites *et al*, 2006).

Due to the paucity of GT-B enzymes whose structures have been solved in the presence of an acceptor, less is known about the mechanisms of GT-B enzymes. Similarly to GT-A, it is thought that inverting GT-Bs use an S_N2-type mechanism. However, both human POFUT1 and Arabidopsis FUT1, inverting GT-Bs, lack a potential catalytic base, and have been proposed to use an S_N1 mechanism (Lira-Navarrete *et al*, 2011; Rocha *et al*, 2016; Moremen & Haltiwanger, 2019).

Both GT-A and GT-B domains have been frequently observed to form symmetric homodimers in crystal structures (Hashimoto *et al*, 2010; Harrus *et al*, 2018). Although the extent to which crystallography can introduce either false positives or false negatives in this respect is not fully known (Harrus *et al*, 2018), the fact that homodimerisation is frequently shown via other techniques such as electrophoresis and analytical ultracentrifugation confirms that many of these interactions are likely to be real (Gibbons *et al*, 2002; El-Battari *et al*, 2003; Kakuda *et al*, 2004). Interestingly, although it has been shown to have both positive and negative effects on activity in different GTs, homodimerisation does not usually bring the active sites of the protomers into close proximity (though they are usually situated on the same side of the complex as each other) (Harrus *et al*, 2018). Therefore, a precise role for dimerisation is yet to be established.

1.3.2 Cytoplasmic tail—transmembrane helix—stem (CTS) domains of glycosyltransferases

Most commonly, Golgi glycosyltransferases are anchored to the Golgi membrane simply by virtue of a CTS domain with a sole transmembrane helix (exceptions include lipid-modifying GTs with a lipophilic surface or transmembrane helix appendage, fully multi-pass GT-C-fold and cellulose-synthase-related GTs, and signal peptide-containing GTs retained by virtue of interactions with other GTs) (Richmond & Somerville, 2000; Breton *et al*, 2001; Lairson *et al*,

2008; Atmodjo *et al*, 2011; Albesa-Jové *et al*, 2014; Hirata *et al*, 2018). Although our understanding of Golgi targeting lags significantly behind that of the rest of the secretory pathway, it has been determined that CTS domains play a critical role in the localisation of GTs, not only to the Golgi apparatus *per se* but also within sub-Golgi compartments: CTS domains therefore represent an important model for understanding eukaryotic membrane trafficking (Grabenhorst & Conradt, 1999; Tu & Banfield, 2010; Welch & Munro, 2019). The function of the CTS domain is not limited to this role, however—the domain often appears to be involved in (and is sometimes essential for) GT homodimerisation and can be subject to functionally important post-translational modification (Breton *et al*, 2001; Young, 2004). Unfortunately, though unsurprisingly, no GT structure containing any typical element of the CTS domain has been published (Harrus *et al*, 2018); therefore, our understanding of CTS domain function is based almost entirely on experiments involving mutagenesis, truncation, and domain-swapping.

Within CTS domains, the stem portion ('stem domain') is usually predicted to be disordered, and it has been proposed that, as flexible linkers, these stems distance the catalytic domain from the membrane, thereby facilitating access to substrates situated towards the centre of the Golgi lumen (Breton *et al*, 2001; Welch & Munro, 2019). Rarely, the stem will contain secondary structure (such as the predicted coiled coil in the heparan initiator EXTL3) or even a fully globular domain: for instance the carbohydrate-binding domain in the *O*-mannosyl glycan synthesis enzyme POMGnT1 (Kuwabara *et al*, 2016; Awad *et al*, 2018). Nevertheless, certain sequence elements (including conserved cysteines) in putatively disordered stem domains have been shown to be essential for Golgi localisation and/or disulphide-mediated dimerisation of some GTs (Young, 2004; Tu & Banfield, 2010; Becker *et al*, 2018). How such elements are recognised in the Golgi lumen by the localisation machinery is currently unknown.

While the stem lacks structure, the transmembrane portion of the CTS domain is predicted to form an α-helix that sits in the organelle membrane (White & Wimley, 1999; Van Dijk *et al*, 2008). Aside from its role as an anchor, the transmembrane helix (TMH) has been implicated in more active methods of localisation. One of the first mechanisms to be proposed relied on the fact that Golgi GT TMHs are shorter than (and have a distinct amino acid composition compared with) TMHs found in the plasma membrane (Bretscher & Munro, 1993). Indeed, experimental evidence continues to support the possibility that gradients in bilayer thickness and lipid composition across secretory pathway membranes help to partition GTs by the

properties of their transmembrane domains (Parsons *et al*, 2019; Welch & Munro, 2019)—equally, TMHs may actually be recognised by integral membrane proteins that carry out this role, such as the COPI adapter TM9SF (Welch & Munro, 2019).

Other early models for Golgi GT localisation proposed that GTs could be retained by virtue of oligomerisation via the TMH or stem domain, though the relevance of such models has become unclear since the development of the cisternal maturation paradigm (Gleeson *et al*, 1994; Colley, 1997; Tu & Banfield, 2010). Nevertheless, it is now known that the TMHs of at least several Golgi glycosyltransferases form disulphide-linked homodimers, and that in some cases the loss of this interaction disrupts normal localisation (Young, 2004; Tu & Banfield, 2010). In addition, it has been established that cysteine residues inside or close to the cytoplasmic end of Golgi TMHs can be modified by *S*-acylation, which appears to inhibit disulphide-mediated dimerisation in some cases (Roth *et al*, 2006; Chumpen Ramirez *et al*, 2017). *S*-acylation can bring about association with particular lipid domains, affect Golgi retention, and induce helix tilt—which can itself facilitate oligomerisation of TMHs (Charollais & Van Der Goot, 2009).

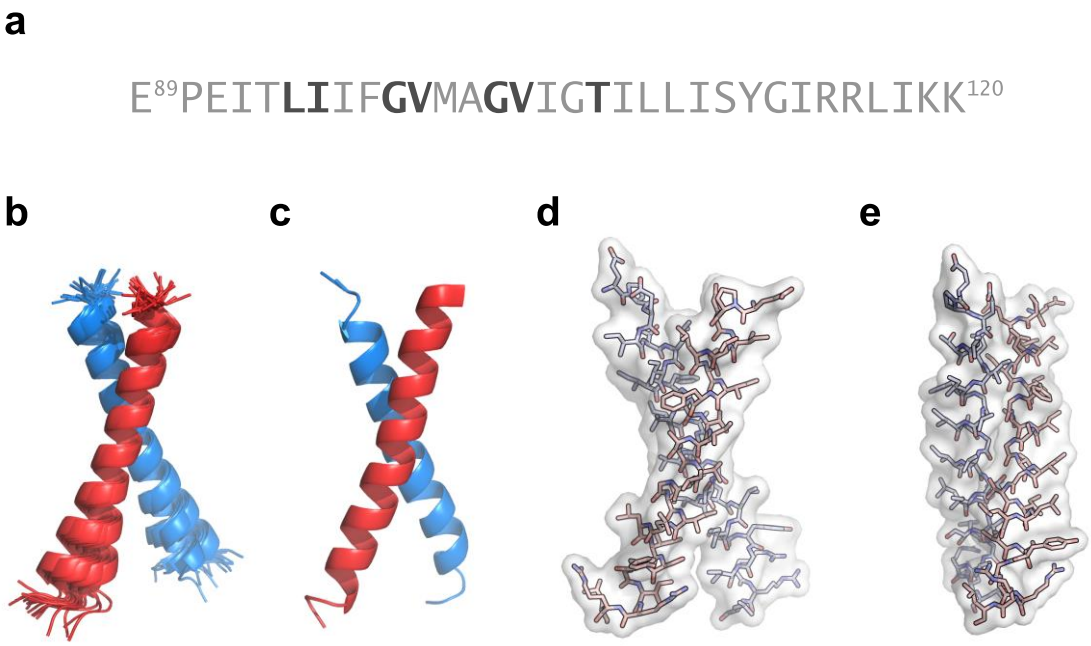


Figure 1.8 The transmembrane helix (TMH) of human erythrocyte protein glycophorin A (GpA) forms a homodimer. a Sequence of the GpA transmembrane region. Residues at the dimer interface are shown in bold. **b** NMR structure (1AFO) of the GpA TMH (representation of the 20 lowest energy states shown). **c** Crystal structure (5EH4) of the GpA TMH. **d** Surface/sticks view of **c**. **e** Rotation of **d**, showing dimer interface.

More widely, computational work has predicted that as many as 29% of all single-pass transmembrane helices in the Golgi exist as homodimers (Pogozheva & Lomize, 2018). Indeed, single-pass TMH dimerisation has been studied in detail over the past three decades, often in relation to transmembrane sequence motifs. Crystallographic and NMR structural data (for TMH pairs in multi-pass membrane proteins and for single-pass TMH dimers, respectively) have revealed two common dimerisation modes: GAS_{right} (a right-handed interaction with a ~40° crossing angle) and GAS_{left} (a left-handed interaction with a ~20° crossing angle), so-named for the prevalence of small amino acids (**g**lycine, **a**lanine, and **s**erine) at the interaction interface (Moore *et al*, 2008). The GAS_{right} mode is often associated with a GxxxG sequence motif, which was originally found in the transmembrane domain of glycophorin A—whose pioneering NMR structure served as a prototype (**Figure 1.8**) (MacKenzie *et al*, 1997; Teese & Langosch, 2015). In such structures, the GxxxG creates a shallow groove on one side of each helix that maximises van der Waals interactions and permits C_{α}–H hydrogen bonding between the two chains (Moore *et al*, 2008; Teese & Langosch, 2015; Anderson *et al*, 2017). The motif has since been generalised to SmxxxSm, where Sm is a small amino acid; tandem repeats of the motif (such as GxxxGxxxG) are also common and have been referred to as 'glycine zippers' (Moore *et al*, 2008; Li *et al*, 2012; Teese & Langosch, 2015). Motifs including polar side chains have also been implicated in TMH self-interaction (Moore *et al*, 2008).

GxxxG, SmxxxSm, and glycine zipper motifs are statistically overrepresented in transmembrane helix sequences (Senes *et al*, 2000; Kim *et al*, 2005). However, it has been cautioned that the presence of any such motif does not automatically indicate its involvement in transmembrane interactions—in fact, the functionality of these motifs seems strongly dependent on sequence context and membrane environment (Li *et al*, 2012; Teese & Langosch, 2015; Morise *et al*, 2020). Moreover, TMH association may in fact be mediated (at least to some extent) by lipophobic effects that favour protein-protein over protein-lipid interactions: a mechanism not directly linked to a specific sequence motif (Sparr *et al*, 2005; Li *et al*, 2012). In addition, many TMH dimer structures do not reveal involvement of specific motifs (Li *et al*, 2012; Steindorf & Schneider, 2016). Nevertheless, it seems that complementary helix surfaces, including (but not limited to) GxxxG-type grooves, are overrepresented in random library screens for TMH homo- and heterodimerisation (Russ & Engelman, 2000; Steindorf & Schneider, 2016). Interestingly, small and polar residues in Golgi glycosyltransferase TMHs have been implicated in localisation and TMH dimerisation (Tu & Banfield, 2010; Schoberer *et al*, 2019b); furthermore, glycosyltransferases have featured in a large computational

prediction screen for interface-driven TMH dimerisation (Anderson *et al*, 2017). In spite of this, it appears that GAS_{right} / GxxxG-type motifs have never been thoroughly discussed in the context of Golgi GT dimerisation or localisation (although a brief mention has been made with regards to the ER GT XXYLT1 (Sethi *et al*, 2012)).

Although almost all Golgi GTs possess TMHs, it is thought that the majority spend some part of their lifetime as soluble proteins in the extracellular space. GTs are frequently cleaved at positions in the transmembrane domain, or nearby in the stem; the signal peptide peptidase-related enzyme SPPL3 has been shown to have some responsibility for this in animals (Young, 2004; Stanley, 2011; Welch & Munro, 2019). The function of this phenomenon is yet to be determined, but it seems that some GTs may retain their activity in their secreted form (Young, 2004).

In recent years, much progress has been made in understanding the role of the cytoplasmic tail at the N-terminus of Golgi GTs. Although typically small in size, this domain is naturally the only part of the GT able to interact directly with cytoplasmic factors. Hence, the tail domain has been the focus of experiments investigating the recruitment of trafficking proteins such as COPI, which mediates retrograde transport through the secretory pathway (Beck *et al*, 2009).

A major breakthrough in this field came with the identification of Vps74, a tetrameric protein that can simultaneously bind to COPI and a (F/L)(L/I/V)XX(R/K) consensus motif in the cytoplasmic tails of several yeast Golgi GTs (thereby facilitating their rescue from post-Golgi compartments) (Tu *et al*, 2008; Schmitz *et al*, 2008; Welch & Munro, 2019). Two functional orthologues were found in animals (GOLPH3 and GOLPH3L), but the number of GTs retained by these proteins appears to be low, and the family has no homologue in plants; therefore, other pathways must exist for Golgi GT retention (Welch & Munro, 2019). Indeed, consistent with the 'positive inside' rule, many Golgi GTs possess basic amino acids in their cytoplasmic tails, and roles have been suggested for these amino acids in Vps74/GOLPH3-independent retention mechanisms: both di-arginine motifs (such as RXR) and recently a Φ[K/R]XLX[K/R] motif have been shown to bind directly to COPI in GT cytoplasmic tails (Tu & Banfield, 2010; Liu *et al*, 2018; Welch & Munro, 2019). Though di-arginine motifs are canonically associated with

COPI-mediated Golgi→ER transport, the distinction between di-arginine motifs for ER retention and those for Golgi retention may lie in their proximity to the TMH, with ER retention motifs situated more distally to the membrane (Welch & Munro, 2019). In addition, a non-basic LPYS motif was recently shown to be necessary for Golgi retention of the glycan-processing enzyme MNS3 in plants (Schoberer *et al*, 2019a).

See **Figure 1.9** for an illustration of the Golgi GT homodimer paradigm.

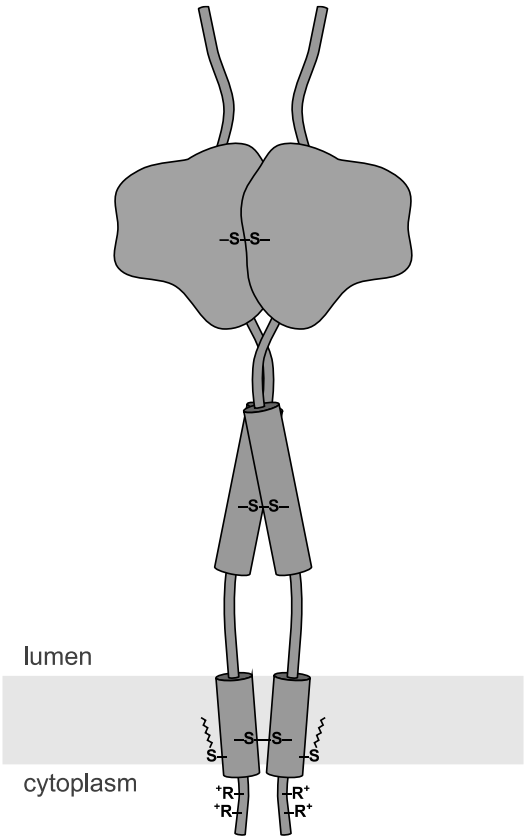


Figure 1.9 An illustration of what some Golgi glycosyltransferase homodimers could look like.

1.3.3 Heteromeric interactions of glycosyltransferases

In addition to the homodimeric self-interaction discussed above, based on co-immunoprecipitation, ER re-localisation, and fluorescent tag-based experiments, GTs have been frequently reported to form hetero-oligomers in the Golgi (Young, 2004; Kellokumpu *et al*, 2016). Tantalisingly, these heteromeric interactions are mostly between GTs involved in the same pathway as each other—the *N*-glycosylation pathway contains various such complexes in animals, plants, and yeast for instance (Young, 2004; Stanley, 2011; Kellokumpu *et al*, 2016). In some cases, hetero-oligomerisation can have dramatic effects on activity: the synergistic activity of the complexed heparan synthases EXT1 (exostosin-1) and EXT2 is a classic example for instance, whereas the uncatalytic GT31-related Cosmc protein is thought to act as a chaperone for T antigen galactosyltransferase C1GalT1 (Wang *et al*, 2010a; Busse-Wicher *et al*, 2014). The potential functions for the formation of such complexes could therefore range from a simple method for increasing local substrate concentrations to a system for building sophisticated, processive machines. Support, at least for the former, lies in the fact that interactions have also been observed between GTs and nucleotide sugar transporters (NSTs), which supply Golgi GTs with their donor substrate (Khoder-Agha *et al*, 2019b). Furthermore, sequence analysis and mapping experiments have revealed that enzymes involved in the synthesis of alternating polysaccharides heparan (*e.g.* EXT1), chondroitin (*e.g.*

CHSY1), and matriglycan (*e.g.* LARGE1) contain two catalytic glycosyltransferase domains, responsible for the two types of linkage in their product respectively (Campbell *et al*, 1997; Kitagawa *et al*, 2001; Coutinho *et al*, 2003; Inamori *et al*, 2012; Busse-Wicher *et al*, 2014; Chen *et al*, 2018). It remains to be determined whether substrates could be channelled between these two domains.

Meanwhile, cryo-electron tomography has revealed that, in the Golgi stacks of the green alga *Chlamydomonas reinhardtii*, arrays of proteins on opposing cisternal membranes may interact with each other across the Golgi lumen in a 'zipper-like' fashion ('trans' interaction) (Engel *et al*, 2015). However, from the results of Förster resonance energy transfer (FRET) and bimolecular fluorescence complementation (BiFC) experiments (based on fluorescent reporter proximity), it appears that hetero-oligomerisation can be mediated by interactions in either the catalytic domain *or* the CTS domain of Golgi GTs (Schoberer *et al*, 2013; Kellokumpu *et al*, 2016). In addition, based on further FRET experiments involving stem truncations, it has recently been suggested that heteromeric interactions of catalytic domains can occur laterally in the same membrane ('cis' interaction), though the possibility for 'trans' interaction was not excluded (Khoder-Agha *et al*, 2019a).

In a landmark comprehensive study of protein complexes, it was observed that non-bijective complexes, that is, complexes in which identical protomers do not also form identical interactions with the rest of the complex (**Figure 1.10**), are rare in nature (Ahnert *et al*, 2015). However, since many of the GTs that are reported to exist as hetero-oligomers have been shown to simultaneously form symmetric, disulphide-linked homodimers (by techniques such as crystallography, electrophoresis, and FRET itself), it would seem that, without significant flexibility of the stem domain, lateral (cis-type) hetero-oligomerisation of these GTs would require the formation of a non-bijective complex, in spite of this rule. Alternatively, it has been suggested that homodimerisation is reversible and is replaced by heterodimerisation as GTs progress through the Golgi, with concomitant remodelling of intermolecular disulphide linkages (Kellokumpu *et al*, 2016; Hassinen *et al*, 2019). Nevertheless, it appears that heteromeric interactions are necessary earlier in the secretory pathway for the transport of several GTs from the ER to the Golgi (McCormick *et al*, 2000; Atmodjo *et al*, 2011; Jiang *et al*, 2016; Zeng *et al*, 2016). How this phenomenon can be reconciled with our understanding of typical protein complex formation will require further work: so far, not a single heteromeric GT–GT complex has been structurally characterised (Khoder-Agha *et al*, 2019a). Given their

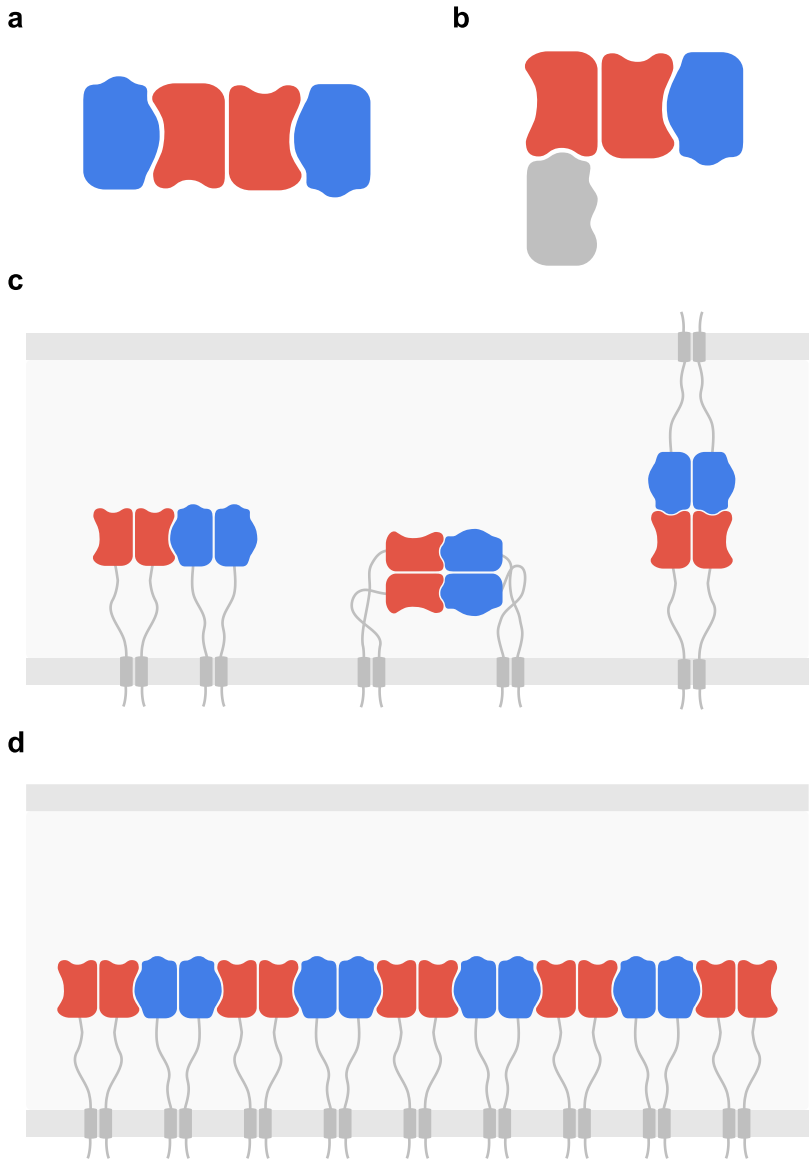


Figure 1.10 Symmetry considerations for heteromeric interaction of homodimeric glycosyltransferases. a Example of a bijective complex (both red and blue subunits experience the same interactions). **b** Example of a non-bijective complex. **c** (left to right) Non-bijective *cis*-interaction of two GT homodimers, bijective *cis*-interaction requiring stem flexibility, and bijective *trans*-interaction. **d** *Cis* interactions could also maintain bijectivity through formation of large protein oligomers or polymers.

ampleness in unstructured domains and loops (Breton *et al*, 2006), it is possible that GTs may not utilise conventional interaction surfaces in their heteromeric interactions.

1.4 CAZy families GT43, GT47, and GT64

Despite similarity in their sequence and structure, GTs in the same CAZy family frequently possess diverse and divergent activities. Therefore, aside from their stereoselectivity, related

GTs from different kingdoms often have vastly disparate functions. In this thesis, the CAZy families GT43, GT47, and GT64 have been selected for investigation: not only because their predicted activities differ substantially between plants and animals, but also because they produce some of the most functionally important polysaccharides in eukaryotes. By analysing GT families from such a perspective, it is hoped that we can obtain a broader understanding of glycosyltransferase biology, including such aspects as evolution, membrane trafficking, and substrate specificity, as well as a means to predict, engineer, and synthesise novel glycan structures and GT activities.

1.4.1 GT43 glycosyltransferases

CAZy family GT43 is a family of inverting GT-A-fold glycosyltransferases unique to eukaryotes (Anders & Dupree, 2011; Taujale & Yin, 2015). Although it was initially thought that GT43s are unique to plants and animals (Fondeur-Gelinotte *et al*, 2006), GT43 homologues have since been found in fungi, chlorophyte algæ, and protists (Taujale & Yin, 2015). Most animal and plant genomes encode at least three or four GT43s (Anders & Dupree, 2011).

In humans, three GT43 enzymes have been identified: GlcAT-P (encoded by *B3GAT1*), GlcAT-S (*B3GAT2*), and GlcAT-I (*B3GAT3*) (Anders & Dupree, 2011). All three are β1,3-glucuronosyltransferases to a terminal galactose (i.e. they transfer β-glucuronic acid to the galactose C3 hydroxyl); however, their acceptor specificities differ. GlcAT-P and GlcAT-S are both active on acceptors with terminal *N*-acetyllactosamine (Gal-β1,4-GlcNAc) (**Figure 1.11**), and are thus involved in the synthesis of the HNK-1 epitope (HSO_3-3-GlcA-β1,3-Gal-β1,4-GlcNAc-…), so-named because it was originally thought to be specific to human natural killer cells (Yamamoto & Oka, 2001). GlcAT-S is also active on acceptors with terminal Gal-β1,3-GlcNAc, and seems to have highest activity on substrates in *N*-linked glycans (Morita *et al*, 2008). On the other hand, GlcAT-I transfers GlcA to the Gal-β1,3-Gal-β1,4-Xyl-α1-Ser moiety in nascent proteoglycans, thus completing the GAG linkage tetrasaccharide (Ouzzine *et al*, 2000). GlcAT-I is therefore critical for the subsequent synthesis of heparan sulphate, chondroitin sulphate, and dermatan sulphate (see *Section 1.5.1* below for a description of heparan sulphate synthesis). Closely related GTs and activities have been characterised in rat, mouse, *Drosophila melanogaster*, and *Caenorhabditis elegans* (Anders & Dupree, 2011).

The structures of GlcAT-P, -S, and -I have each been solved by X-ray crystallography (Pedersen *et al*, 2000, 2002; Kakuda *et al*, 2004; Shiba *et al*, 2006). All three structures constitute catalytic-domain homodimers with identical interaction surfaces, and all exhibit

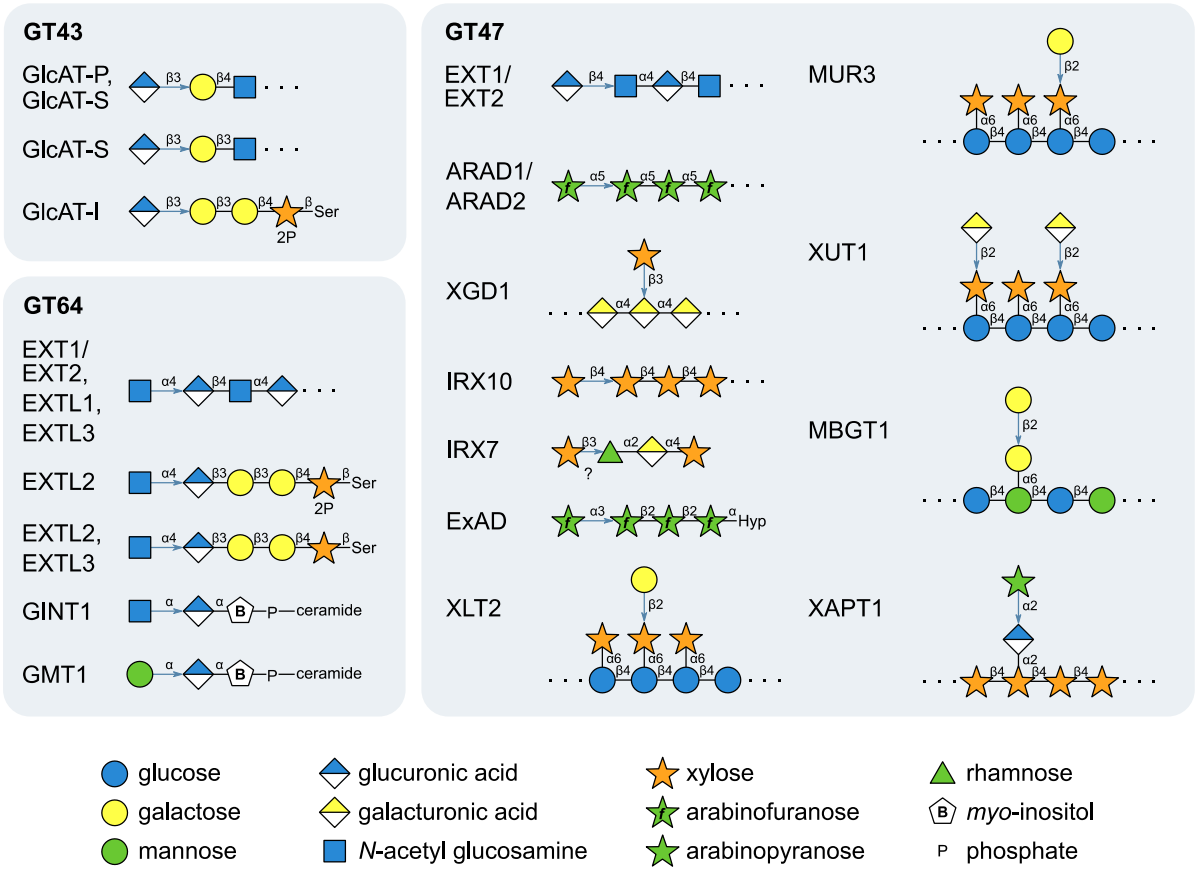


Figure 1.11 Reported and speculative activities of GT43, GT47, and GT64 glycosyltransferases from *Homo sapiens* and *Arabidopsis thaliana*.

amino acid side chains from one protomer extending into the active site of the other. Homodimerisation of GlcAT-P has been independently confirmed by analytical centrifugation, while it has also been shown that GlcAT-I homodimers are stabilised in the stem region by an intermolecular disulphide bridge at Cys-33 (Ouzzine *et al*, 2000). Furthermore, these structures have inspired several point mutation experiments investigating substrate specificity. For example, the substitution of acceptor-binding residues Trp-234 and Ala-309 in GlcAT-S with their analogues in GlcAT-P (Phe and Val) results in the abrogation of its distinctive Gal-β1,3-GlcNAc glucuronosylating activity (Shiba *et al*, 2006). In addition, following the publication of the first GlcAT-I structure (Pedersen *et al*, 2000), Ouzzine *et al* (2002) investigated the role of two residues in UDP-GlcA binding: His-308 and Arg-277. Interestingly, mutation of His-308 to Arg resulted in a loss of donor specificity, with UDP-Glc, UDP-GlcNac, and UDP-Man (not normally present in animals) becoming potential substrates. This histidine was therefore touted as the determinant of donor substrate specificity in these enzymes (Anders & Dupree, 2011). However, a contemporaneously published structure of GlcAT-I bound with UDP-GlcA

clearly shows that His-308 makes interactions with only the C2 hydroxyl of GlcA (which is invariant between Glc and GlcA) and possibly the Mn^{2+} ion (this residue would today be recognised as the canonical C-His), rather than the distinguishing C6 carboxy group of the normal donor sugar (Pedersen *et al*, 2002). These results indicate the subtlety of glycosyltransferase substrate specificity, as well as the difficulty of its prediction.

Based on their sequence similarity, plant GT43s can be divided into two main groups that diverged from each other before the time of terrestrial colonisation (Taujale & Yin, 2015). Arabidopsis possesses two genes in each: *IRX9* and *IRX9L*, which themselves diverged from one another at around the time of the gymnosperm–angiosperm divergence (see **Section 1.6** for descriptions of taxonomical terms), and *IRX14* and *IRX14L*, which diverged from one another much more recently (Taujale & Yin, 2015). Although the corresponding proteins have not been structurally characterised, BiFC experiments support the notion that they each form homodimers *in planta* (Zeng *et al*, 2016; Jiang *et al*, 2016).

Collectively, the proteins encoded by these genes are essential for the synthesis of the xylan backbone, and the removal of either pair by knockout mutation results in severely reduced xylan content, as well as the concomitant collapsed vessel phenotype that lends them its name: 'irregular xylem' (IRX) (Wu *et al*, 2010a; Rennie & Scheller, 2014). IRX9 and IRX9L appear to be important for xylan synthesis during secondary and primary cell wall production respectively (Mortimer *et al*, 2015; Ratke *et al*, 2015). Surprisingly, however, the precise functions of these proteins are yet to be determined (Smith *et al*, 2017). Curiously, mutation of the normally critical metal-binding DxD motif in IRX9L does not impede its ability to complement the *irx9* mutant phenotype in overexpression experiments, while the DxD motif in IRX9 itself has been lost completely, suggesting that these proteins lack catalytic activity (Ren *et al*, 2014). Supporting this idea are the facts that IRX9 has lost many other residues that are conserved in active GT43s, and that generally, sequences in the IRX9 subgroup have seen a much faster rate of evolution compared with other groups (Taujale & Yin, 2015). In contrast, the DxD motif of IRX14, which is involved in both primary and secondary cell wall xylan synthesis, appears to be essential for its function (Ren *et al*, 2014; Mortimer *et al*, 2015). However, any activity for IRX14 or IRX14L has yet to be demonstrated *in vitro* (Rennie & Scheller, 2014).

The function of these proteins is further confounded by the existence of a third group of factors also required for xylan backbone synthesis: IRX10 and IRX10L. These two enzymes belong

to CAZy family GT47, and have been convincingly shown to possess the β1,4-xylosyltransferase activity required to form the xylan backbone (Jensen *et al*, 2014; Urbanowicz *et al*, 2014). The co-operation of IRX9/9L, IRX14/14L, and IRX10/10L in xylan backbone synthesis will be discussed below in *Section 1.5.2*, but for now, we have arrived at the topic of the GT47 family, and it seems an appropriate time to discuss them.

1.4.2 GT47 glycosyltransferases

The GT47 family comprises a large group of inverting glycosyltransferases found almost exclusively in eukaryotes (Geshi *et al*, 2011; Tan *et al*, 2018). While animal genomes tend to encode two to four GT47 members (Geshi *et al*, 2011; Busse-Wicher *et al*, 2014), the complement of GT47s in a typical plant genome is much larger—with 39 members in Arabidopsis (Xu *et al*, 2018). GT47s have been predicted to adopt a GT-B fold (Liu & Mushegian, 2003; Awad *et al*, 2018); however, despite their prevalence in plant glycan synthesis, no structural data has yet been published to confirm this.

In animals, only one GT47-catalysed activity is known: the β-glucuronosyltransferase activity required for heparan sulphate backbone synthesis (Wei *et al*, 2000; Geshi *et al*, 2011; Busse-Wicher *et al*, 2014) (**Figure 1.11**). Remarkably, the proteins that catalyse this reaction, known as 'exostosins', possess two GT domains, with an N-terminal GT47 domain and a C-terminal GT64 domain that usually carries α1,4-*N*-acetylglucosaminyltransferase activity (Edvardsson *et al*, 2011; Geshi *et al*, 2011). Humans possess five exostosins—EXT1, EXT2, EXTL1, EXTL2, and EXTL3 (though EXTL2 lacks a GT47 domain) (Busse-Wicher *et al*, 2014; Chen *et al*, 2018). Their name derives from the fact that defects in EXT1 and EXT2 cause hereditary multiple exostoses (HME), a genetic condition characterised by multiple benign cartilaginous tumours (D'Arienzo *et al*, 2019). Orthologues of these enzymes have been characterised throughout the animal kingdom, and are often essential for development and/or viability (Busse-Wicher *et al*, 2014). The precise role of exostosin enzymes in heparan synthesis will be discussed below in *Section 1.5.1*.

In contrast, plant GT47s are involved in the synthesis of diverse cell wall polysaccharides and glycans, and have been grouped into six clades based on their sequence (Li *et al*, 2004; Geshi *et al*, 2011). GT47 clade A (GT47-A) is perhaps the best characterised of these; in Arabidopsis, it comprises eleven members, including two xyloglucan β1,2-galactosyltransferases (MUR3 and XLT2) and a xyloglucan β1,2-galacturonosyltransferase (XUT1; see *Section 1.5.3*) (Geshi *et al*, 2011; Pauly & Keegstra, 2016). Recent data has shown that two previously

uncharacterised Arabidopsis members encode a glucomannan β1,2-galactosyltransferase (MBGT1) and a (glucurono)xylan α1,2-arabinopyranosyltransferase (XAPT1) (Yu *et al*, 2021a, 2021b). Furthermore, two xyloglucan α1,2-arabinofuranosyltransferases from tomato (*Sl*XST1 and *Sl*XST2) and a xyloglucan α1,2-arabinopyranosyltransferase from the moss *Physcomitrium* (formerly *Physcomitrella*) *patens* (*Pp*XDT) have been characterised and assigned to this clade as homologues of XLT2 (Schultink *et al*, 2013; Zhu *et al*, 2018). Therefore, the activities observed in GT47-A so far have been limited to a single type of linkage: β1,2 (or α1,2 in the case of L-sugars—see footnote 2 on *p*. 2).

GT47-B has been less extensively characterised. In Arabidopsis, it contains eight members, including a pair of related enzymes that may possess arabinan α1,5-arabinofuranosyltransferase activity and that may form a complex with each other (ARAD1 and ARAD2) (Harholt *et al*, 2006, 2012). A third (uncharacterised) member, AT3G45400, is less closely related but may constitute an orthologue of another putative arabinan α1,5-arabinofuranosyltransferase from *Nicotiana alata* (*Na*ARADL1) (Lampugnani *et al*, 2016), suggesting that this type of activity could be widespread in the clade.

GT47-C is the largest clade in Arabidopsis, with fourteen members. However, so far only one has been characterised: the xylogalacturonan β1,3 xylosyltransferase (XGD1) (Jensen *et al*, 2008). This activity is so far unique amongst characterised GT47s in that the acceptor is not a terminal sugar, but rather a non-terminal residue in the galacturonan backbone itself.

GT47-D contains the previously mentioned IRX10 and IRX10L, for which evidence of xylan β1,4-xylosyltransferase activity exists (Jensen *et al*, 2014; Urbanowicz *et al*, 2014). However, two further Arabidopsis GTs are found in this clade: IRX7/FRA8 and IRX7L/F8H. Knockout of the former leads to a collapsed xylem phenotype, a decrease in xylan content, and the loss of the xylan reducing end oligosaccharide (REO) structure (Xyl-β1,3-Rha-α1,2-GalA-α1,4-Xyl), while overexpression of the latter can fully rescue the *irx7* mutant (Zhong *et al*, 2005; Brown *et al*, 2007; Peña *et al*, 2007; Lee *et al*, 2009). The exact activity of these two enzymes remains to be determined, but it has been suggested that they possess the β1,3-xylosyltransferase activity or possibly the α1,2-L-rhamnosyltransferase activity required for REO synthesis (Peña *et al*, 2007; Scheller & Ulvskov, 2010).

GT47-E and GT47-F comprise only one member each in Arabidopsis: AT3G57630 and AT1G21480, respectively (Li *et al*, 2004). Whereas the activity of AT1G21480 has not been determined, AT3G57630, or ExAD, encodes an α1,3-arabinofuranosyltransferase that

specifically adds the fourth Ara*f* residue in extensin arabinans (Møller *et al*, 2017). GT47 sequences from chlorophyte algæ are consistently grouped in the ExAD-related clade, which has led to suggestions that this clade is ancestral to other plant GT47s (Geshi *et al*, 2011; Ulvskov *et al*, 2013; Møller *et al*, 2017; Xu *et al*, 2018).

1.4.3 GT64 glycosyltransferases

GT64-family glycosyltransferases adopt the GT-A fold and employ a retaining mechanism of activity. Although an earlier structural analysis did not reveal the similarity (Hashimoto *et al*, 2010), a recent phylogenetic analysis has suggested that GT64 is the most closely related family to inverting family GT43 (Taujale *et al*, 2020). Interestingly, although GT64s use a retaining mechanism, they nevertheless maintain a conserved aspartate residue at the position at which the catalytic base of an inverting enzyme would be expected; quantum mechanical/molecular mechanical simulations using the GT64 domain of EXTL2 indicate that this residue in fact helps to stabilise the oxocarbenium ion intermediate (Mendoza *et al*, 2017). It has been proposed that the retaining mechanism of GT64s has arisen from a spatial perturbation of the ancestral catalytic base (which is present as glutamate in GT43 enzymes) (Mendoza *et al*, 2016; Taujale *et al*, 2020).

Animal GT64s are entirely represented by the (mostly) bi-domain exostosins introduced above. Typically, these GT64 domains possess heparan α1,4-*N*-acetylglucosaminyltransferase activity (Edvardsson *et al*, 2011), though the human enzyme EXTL2 appears to possess a distinct acceptor specificity (see *Section 1.5.1*) (Kitagawa, 2019). Furthermore, in terms of its *donor* specificity, EXTL2 has been shown to possess an additional α1,4-*N*-acetylgalactosaminyltransferase activity, though a biological purpose for this activity is yet to be discovered (Kitagawa *et al*, 1999; Kitagawa, 2019). Perhaps thanks to its lack of a GT47 domain, EXTL2 has been successfully crystallised; its structure has been solved both with UDP-GlcNAc and with UDP-GalNAc bound (Pedersen *et al*, 2003). Additionally, isothermal titration calorimetry has shown that EXTL2 may also bind UDP-GlcA, though it is not thought to hydrolyse this molecule (Sobhany *et al*, 2005).

The functions of plant GT64s have only been brought to light in the past four years. Arabidopsis possesses three GT64s: GMT1/EPC1, GINT1, and a remaining uncharacterised enzyme AT1G80290, which lacks a transmembrane domain (Edvardsson *et al*, 2011). GMT1 and GINT1 are both involved in the synthesis of GIPC glycolipids—more specifically, in the addition of the second sugar moiety. GMT1 adds mannose to the GlcA-IPC acceptor, whereas

GINT1 appears to add *N*-acetylglucosamine instead (Fang *et al*, 2016; Ishikawa *et al*, 2018) (**Figure 1.11**). Mutation of GMT1 results in a severe stunted-growth phenotype (Edvardsson *et al*, 2011).

1.5 Biosynthesis of heparan, xylan, and xyloglucan

1.5.1 Heparan sulphate synthesis

Owing to the importance of glycosaminoglycans in cell–cell communication and growth regulation, the synthesis of heparan sulphate (HS), chondroitin sulphate (CS), and their derivatives (heparin and dermatan sulphate) is tightly regulated (Couchman & Pataki, 2012; Prydz, 2015). The process always begins in the same way: with the formation of a 'GAG linkage region', which consists of a GlcA-β1,3-Gal-β1,3-Gal-β1,4-Xyl tetrasaccharide attached to a serine residue in the (variable) proteoglycan backbone (**Figure 1.12**) (Esko *et al*, 2009). Synthesis of the linkage region commences with serine xylosylation by XylT1 or XylT2, typically at an SG motif (owing to the narrow peptide binding cleft of these enzymes), and is followed by galactosylation by GalT-I (Kreuger & Kjellén, 2012; Briggs & Hohenester, 2018). At this disaccharide stage, the xylose may be 2-*O*-phosphorylated by the kinase FAM20B, stimulating the final two steps: galactosylation by GalT-II followed by glucuronosylation by the GT43 member GlcAT-I (Tone *et al*, 2008; Koike *et al*, 2009; Wen *et al*, 2014; Zhang *et al*, 2018b). This last step is normally accompanied by rapid dephosphorylation by the phosphatase XYLP, which is necessary for the progression of GAG synthesis (Nadanaka *et al*, 2013; Koike *et al*, 2014). However, it appears that dephosphorylation can be blocked by EXTL2—this enzyme is currently thought to add α-GlcNAc preferentially to the phosphorylated linkage region, thereby terminating GAG synthesis at this stage (Kitagawa, 2019).

Despite the fact that many proteoglycans show preferences as to whether they are primarily modified with either HS or CS (Noborn *et al*, 2016), exactly how the Golgi glycosylation machinery decides whether to subsequently add α-GlcNAc (for HS) or β-GalNac (for CS) to the linkage region is not yet fully understood. Nevertheless, some patterns have been observed. For example, HS sites seem to be enriched in multiple acidic amino acids (though acidic amino acids are also frequently found at CS sites, and are generally preferred at the −4 to −2 positions by XylT-1), one or more hydrophobic residues, and repeats of the SG motif; in addition, glypican-1, which has an unusually strong preference for HS, appears to possess more distal sequence elements underlying its predisposition (Esko & Zhang, 1996; Esko & Selleck, 2002; Prydz, 2015). Since the HS synthetic machinery is localised in the *cis*-Golgi, whereas the CS machinery is found in *trans*-Golgi (*i.e.* later in the secretory pathway), it has been suggested

that addition of CS is the default in the absence of specific HS addition motifs (Esko & Zhang, 1996; Fransson *et al*, 2000; Kreuger & Kjellén, 2012; Prydz, 2015). Nevertheless, the linkage region of CS glycans alone is often found to contain 4-*O*- or 6-*O*-sulphatated galactosyl residues, and it has been suggested that such modifications may block HS synthesis (Esko & Selleck, 2002; Prydz, 2015). However, ectopic expression of CS 6-*O*-sulphotransferase 1 (which can sulphate the linkage region *in vitro*) in Chinese hamster ovary (CHO) cells does not affect the HS/CS ratio (Chen *et al*, 2018).

The first committed step in HS synthesis is the addition of the very first α-GlcNAc to the linkage tetrasaccharide (GlcNAcT-I activity) (Kitagawa & Nadanaka, 2002; Esko *et al*, 2009).

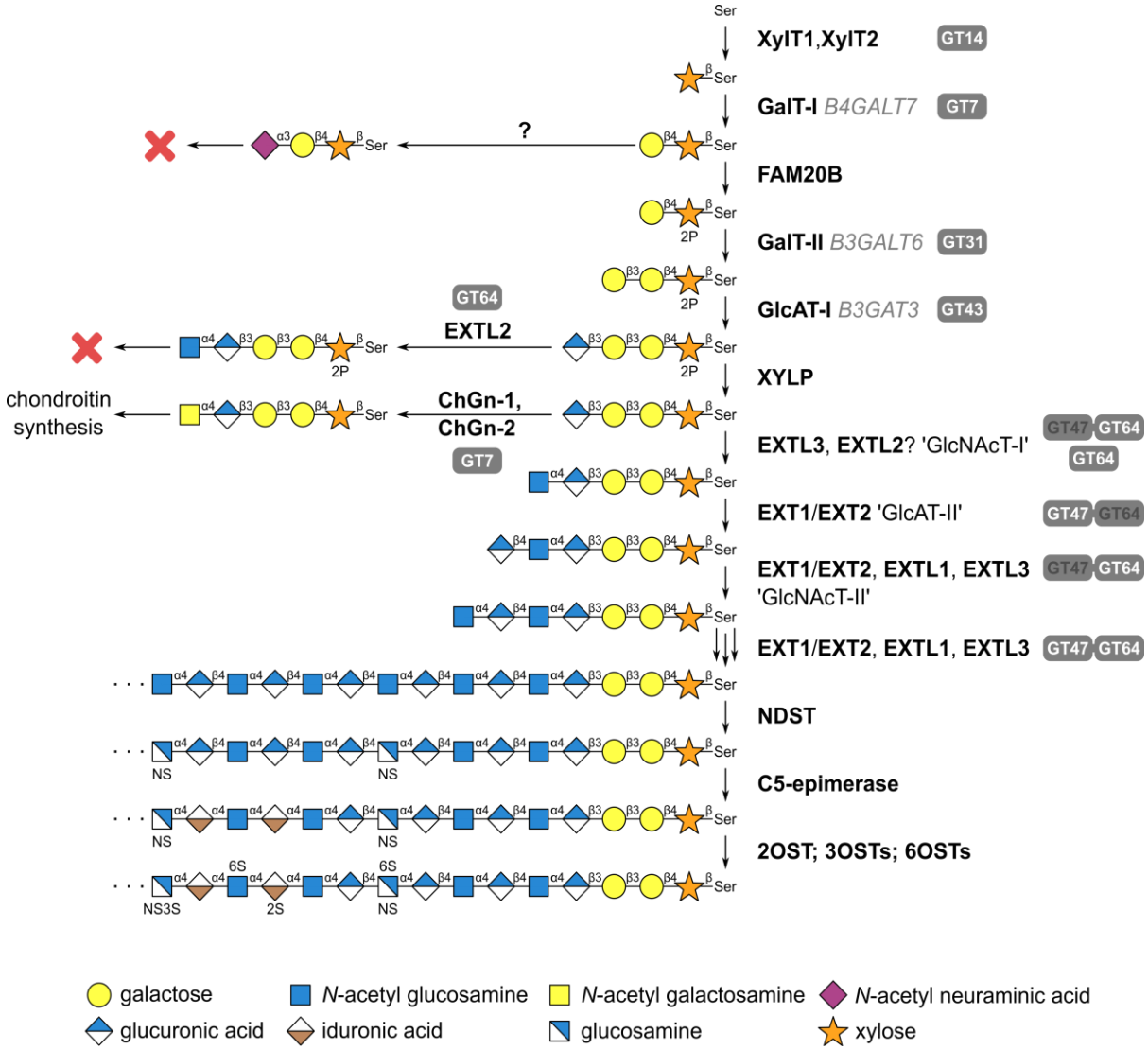


Figure 1.12 Synthesis of heparan sulphate in *Homo sapiens*. Labels in grey boxes indicate CAZy GT families.

Whilst EXTL2 has been shown to have some α1,4-*N*-acetylglucosaminyltransferase activity

towards an unphosphorylated acceptor analogue *in vitro*, this protein is found only in vertebrates, and its activity seems to be generally much less important than that of EXTL3, probably the only other human exostosin to exhibit this activity (though EXT1/EXT2 appear able to initiate HS chains from an acceptor analogue with a hydrophobic aglycone) (Kitagawa *et al*, 1999; Kim *et al*, 2003; Busse-Wicher *et al*, 2014). Consistent with the role of EXTL3 as the sole (or most important) initiator of HS backbone synthesis, EXTL1 and EXTL2 are not widespread in animals: in general, exostosins from across the kingdom form three groups corresponding to orthologues of EXT1, EXT2, and EXTL3 respectively (Feta *et al*, 2009; Busse-Wicher *et al*, 2014). In fact, HS-synthesising *C. elegans* possesses merely an EXTL3 homologue (rib-2) and a GT47-only EXT1 homologue (rib-1) (Busse-Wicher *et al*, 2014). Furthermore, knockout of the mouse EXTL3 orthologue results in a complete lack of HS and embryo death after nine days; similarly, knockout of the EXTL3 orthologue in CHO cells results in total abrogation of HS synthesis (Takahashi *et al*, 2009; Chen *et al*, 2018). Mutants in EXTL3 orthologues in zebrafish, *D. melanogaster*, and *C. elegans* also exhibit impaired HS synthesis (Busse-Wicher *et al*, 2014). In human embryonic kidney 293 (HEK293) cells, knockdown of EXTL3 appears to increase chain length (perhaps by limiting the number of chains available for extension), while overexpression has no effect on length (Busse *et al*, 2007). In contrast, knockdown of EXTL3 in human breast cancer cells has been reported to reduce both the amount and length of HS chains (Nadanaka *et al*, 2014). In terms of human disease, deleterious mutations to the *EXTL3* gene can result in a range of developmental and neurological diseases (Paganini *et al*, 2019).

Following initiation, extension of the heparan backbone occurs through the addition of β-GlcA and α-GlcNAc residues in an alternating fashion (GlcAT-II and GlcNAcT-II activities respectively) (Kitagawa & Nadanaka, 2002). EXT1 and EXT2 are the main enzymes responsible for this, and are thought to form a heteromeric complex (Busse-Wicher *et al*, 2014). *In vitro* experiments have shown that, when purified in isolation, EXT1 possesses comparable levels of GlcAT-II and GlcNAcT-II activity, whereas EXT2 possesses a only a very small amount of GlcAT-II activity; in contrast, the co-purified complex has overwhelming GlcAT-II activity (McCormick *et al*, 2000; Busse & Kusche-Gullberg, 2003). EXTL3 and EXTL1 (the latter being expressed in only a few tissues) have also each been shown to possess GlcNAcT-II activity but, in spite of their GT47 domains, neither has been reported to possess GlcAT activity (Kim *et al*, 2001; Busse *et al*, 2007; Busse-Wicher *et al*, 2014).

Concomitantly with extension, the nascent heparan backbone undergoes significant modification. Many GlcNAc residues undergo simultaneous deacetylation and 2-*N*-sulphation by *N*-deacetylase/*N*-sulphotransferases (NDSTs); this is usually followed by the conversion of most GlcA residues to iduronic acid by uronosyl C5 epimerases, the 2-*O*-sulphation of most of those IdoA residues by 2-*O*-sulphotransferases, and 3/6-O-sulphation of many GlcNAc/GlcNS residues (Kreuger & Kjellén, 2012). The unique pattern of backbone modifications and decorations determines the ability of HS to bind a wide range of targets with high specificity (Esko & Selleck, 2002).

Glycosaminoglycan analogues have also been isolated from bacteria. For example, the K5 strain of *E. coli* secretes a polysaccharide of alternating 1,4-linked β-GlcA and α-GlcNAc units (identical in structure to the unmodified heparan backbone) into its extracellular capsule in order to evade detection by the human immune system (Wang *et al*, 2010b). However, the enzymes that synthesise this polysaccharide are completely unrelated to exostosins, and belong to GT2 and GT45 (DeAngelis, 2002).

1.5.2 Xylan synthesis

β1,4-Linked xylans are present in all land plants, and seem to have first arisen in the streptophyte algæ (Hsieh & Harris, 2019). However, compared with heparan, the synthesis of the xylan backbone is far less well understood. Nevertheless, genetics experiments have revealed the GTs involved. In Arabidopsis, it is thought that IRX9L, IRX10L, and IRX14 act together to make xylan during primary cell wall synthesis, while IRX9, IRX10, IRX14/IRX14L make secondary cell wall xylan (though there exists some redundancy within each pair of homologues) (Mortimer *et al*, 2015; Ratke *et al*, 2015; Zhong *et al*, 2018). Numerous functional orthologues of these enzymes have been characterised in a wide range of plants—including an IRX10 orthologue from the streptophyte alga *Klebsormidium nitens* (Zhou *et al*, 2007; Lee *et al*, 2011; Chen *et al*, 2013; Lovegrove *et al*, 2013; Jensen *et al*, 2014, 2018; Ratke *et al*, 2018; Pellny *et al*, 2020; Petrik *et al*, 2020).

Despite the fact that the xylan backbone contains only one type of glycosidic linkage, and that IRX10/10L alone can catalyse its formation *in vitro* (albeit inefficiently) (Jensen *et al*, 2014; Urbanowicz *et al*, 2014), all three types of GT (IRX9/9L, IRX10/10L, IRX14/14L) are required for xylan backbone synthesis *in vivo* (Wu *et al*, 2009, 2010b; Brown *et al*, 2009; Keppler & Showalter, 2010). This, the lack of IRX9 catalytic activity, and the fact IRX10/10L lacks a transmembrane domain to anchor it to the Golgi (Wu *et al*, 2009) have led to the proposal that

a trimeric complex must be formed: the so-called 'xylan synthase complex' (XSC) (York & O'Neill, 2008; Smith *et al*, 2017). Biochemical evidence for this idea has so far come from two groups of experiments, using IRX orthologues in wheat (*Triticum æstivum*) and asparagus (*Asparagus officinalis*), respectively (Zeng *et al*, 2010; Jiang *et al*, 2016; Zeng *et al*, 2016). However, the reported complex membership and stoichiometry differ strikingly between these studies. In Jiang *et al* (2016), the wheat XSC was proposed to contain dimeric *Ta*GT43-4 (an IRX14 orthologue), monomeric *Ta*GT47-13 (an IRX10 orthologue interacting directly with *Ta*GT43-4), and several other proteins without an immediate connection to xylan backbone synthesis. In contrast, Zeng *et al* (2016) proposed that the asparagus XSC contains homodimeric *Ao*IRX9, homodimeric *Ao*IRX10, homodimeric *Ao*IRX14A, and potentially a further protein bridging *Ao*IRX10 to the GT43s. It seems unlikely that the interaction of so many polypeptides could occur bijectively (see *Section 1.3.3*).

As mentioned in the previous section, the presence of GT47-family enzymes IRX7 and IRX7L have been linked to with the presence of the xylan reducing end oligosaccharide (REO) structure (Xyl-β1,3-Rha-α1,2-GalA-α1,4-Xyl), which has been detected in many plants (though not in grasses, or the lower plants *P. patens* and *Selaginella moellendorffii*) (Scheller & Ulvskov, 2010; Rennie & Scheller, 2014; Smith *et al*, 2017). The predicted GTs IRX8 and PARVUS are also required for its detection *in planta*; however, their precise activities have not been determined (Rennie & Scheller, 2014). It has been suggested that the REO may function as a primer for xylan backbone synthesis (Brown *et al*, 2007; Lee *et al*, 2007). The fact that GT47s and GT43s co-operate in backbone initiation and elongation for both xylan and heparan has been noted previously, and the REO has been compared with the GAG linkage region (Smith *et al*, 2017). However, so far, there has been no evidence to suggest that the presence of these two GT families in both pathways constitutes anything more than convergent evolution. A consistent scaffold for xylan synthesis (such as a core protein) has yet to be identified, but a structure containing an arabinogalactan protein-linked arabinoxylan has been reported in Arabidopsis (Tan *et al*, 2013).

The xylan backbone is decorated with a range of sugars and functional groups depending on tissue and species (Rennie & Scheller, 2014). Very often, xylosyl residues undergo 2/3-*O*-acetylation by TBL-family enzymes (though not in conifers) and 2-*O*-glucuronosylation by GUX enzymes (**Figure 1.13a**; though some grass xylans lack GlcA) (Smith *et al*, 2017). GlcA residues often undergo further 4-*O*-methylation by GXM methyltransferases (Wierzbicki *et al*,

2019). In some xylan chains, the [Me]GlcA and acetate decorations are limited to alternately numbered xylosyl residues; this even spacing permits a two-folded screw conformation that facilitates docking into cellulose microfibrils (Grantham *et al*, 2017). 2/3-*O*-

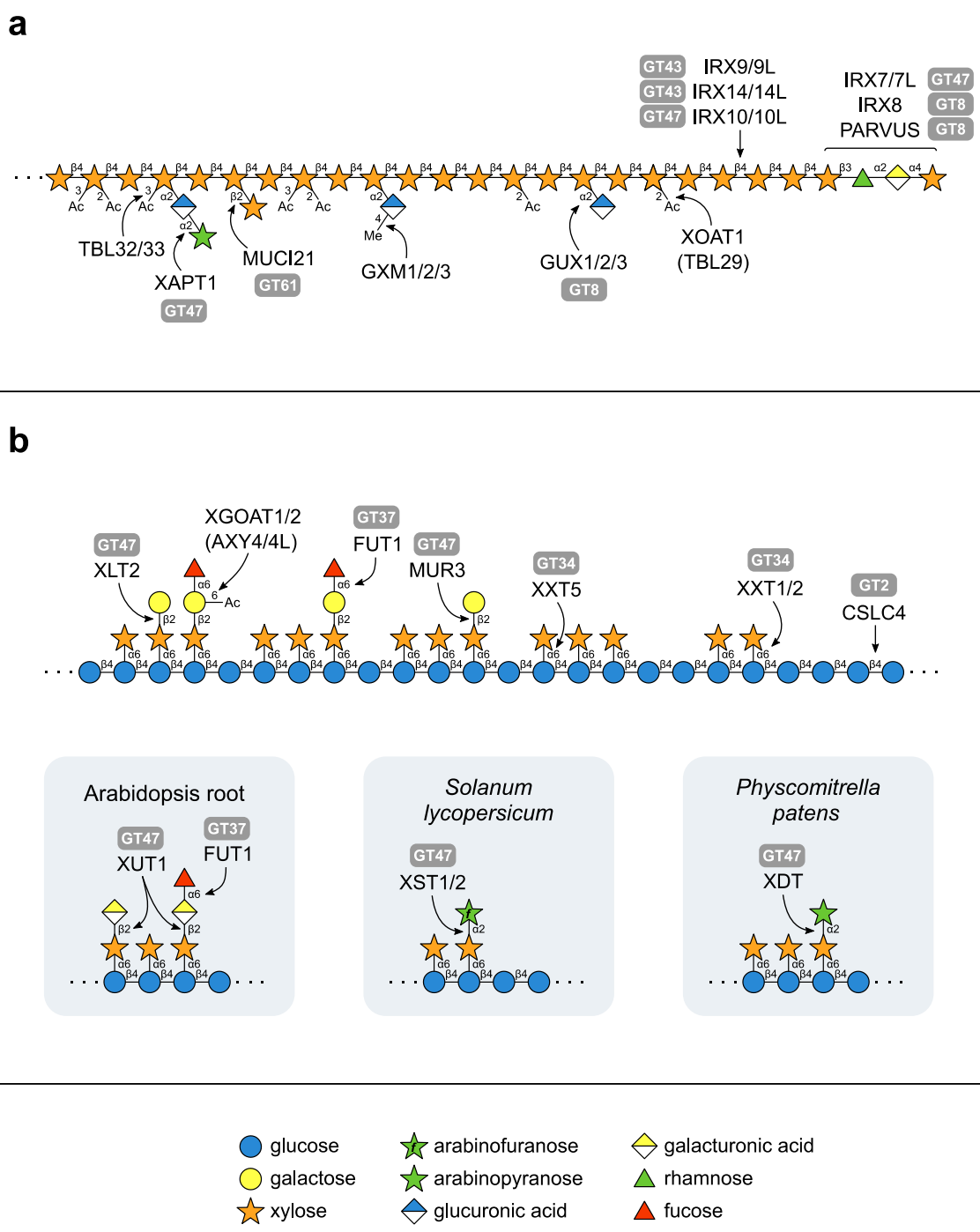


Figure 1.13 Xylan and xyloglucan synthesis in *Arabidopsis thaliana*. a Golgi enzymes involved in the biosynthesis of xylan. **b** Golgi enzymes involved in the biosynthesis of xyloglucan. Labels in small grey boxes indicate CAZy GT families. Large grey boxes show scenarios involving alternative GT47 activities.

arabinofuranosylation and/or 2-*O*-xylosylation of backbone Xyl residues by GT61-family enzymes is also common, especially in grass xylans, where Ara*f* residues can be cross-linked by additional diferulate moieties (Smith *et al*, 2017; Hatfield *et al*, 2017). Cereal grain xylans can be especially complex, with Xyl-Ara*f* and L-Gal-Xyl-Ara*f* sidechains branching from the backbone (Smith *et al*, 2017; Beri *et al*, 2020). Finally, in primary cell wall xylan of non-grass species, $^{[Me]}$GlcA residues can be 2-*O*-arabinopyranosylated, while in eucalyptus, galactose may be added instead (Smith *et al*, 2017). Recent data have shown that both activities are derived from enzymes in clade A of the GT47 family: xylan arabinopyranosyltransferase 1 (XAPT1) and *Eucalyptus grandis* xylan galactopyranosyltransferase (*Eg*XLPT), respectively (Yu *et al*, 2021b).

1.5.3 Xyloglucan synthesis

The synthesis of the plant cell wall polysaccharide xyloglucan (XyG) is much better understood. However, its structure can vary considerably between species and tissues; hence, only a brief summary is given here.

The backbone of xyloglucan is chemically identical to a cellulose chain, consisting entirely of β1,4-linked glucosyl residues (Scheller & Ulvskov, 2010). Therefore, it is not surprising that in Arabidopsis, the XyG backbone is made by enzymes from the CSLC family, which are directly related to cellulose synthase (Liepman & Cavalier, 2012; Kim *et al*, 2020). Whereas CesAs are found in the plasma membrane, CSLC4, the best characterised CSLC member, is thought to sit in the membrane of the Golgi body (Liepman & Cavalier, 2012). Therefore, with its GT2 domain presumably in the cytoplasm, CSLC4 is thought to consume cytosolic UDP-Glc and extrude its glucan product into the Golgi lumen (Pauly & Keegstra, 2016). This glucan is subsequently decorated with α1,6-linked xylosyl residues by XXT1, XXT2, and XXT5; these enzymes have been reported to form a functional complex with CSLC4 (Chou *et al*, 2012; Zabotina, 2012; Culbertson *et al*, 2016). Xylosylation often occurs in a four-residue repeat: in Arabidopsis, usually every fourth glucose is undecorated (**Figure 1.13b**; 'XXXG' by the standard nomenclature (Fry *et al*, 1993)), whereas in some other species, including grasses, tomato, and lower plants, only the first two glucoses in the repeat are typically xylosylated ('XXGG') (Pauly & Keegstra, 2016; Zavyalov *et al*, 2019). XyG xylose sidechains can be 2-*O*-galactosylated: in Arabidopsis primary cell walls this is achieved by GT47-family enzymes XLT2 (for the second Xyl, producing XLXG) and MUR3 (for the third Xyl, producing 'XXLG') (Zabotina, 2012; Pauly *et al*, 2013). The MUR3 knockout mutant *mur3-3* exhibits a severe dwarf phenotype with 'cabbage-like' growth (in contrast to the comparatively weak

phenotypes of the point mutants *mur3-1* and *mur3-2*, despite their apparent lack of third-position galactose, and the knockout mutant *xlt2*) (Tamura *et al*, 2005; Jensen *et al*, 2012; Kong *et al*, 2015). Almost all Gal residues added by MUR3 are thought to be 6-*O*-fucosylated by FUT1 (producing 'XXFG' and 'XLFG')—though mutation of FUT1 does not result in a discernible growth phenotype (Günl *et al*, 2011; Kong *et al*, 2015; Soto *et al*, 2019). Fucosidases later remove some of these fucosyl residues in the cell wall (Pauly & Keegstra, 2016). In liverworts, mosses, and Arabidopsis root, xylosyl residues may instead be 2-*O*-galacturonosylated (in Arabidopsis, this is achieved by GT47-family XUT1), while in ferns and certain higher plants (including tomato), arabinofuranosyl secondary substitutions are common (Pauly & Keegstra, 2016). These Ara*f* residues are added by *Sl*XST1 and *Sl*XST2 in tomato (Schultink *et al*, 2013). In *P. patens*, Ara*p* can be added instead by *Pp*XDT, an enzyme that is closely related to another GT47-A enzyme with XLT2-like activity (*Pp*XLT2) (Zhu *et al*, 2018). In addition to arabinose, the xyloglucan of some plants from the Ericales order (such as cranberry and argan) also exhibit secondary substitutions of β-xylose (Ray *et al*, 2004; Hilz *et al*, 2007; Hotchkiss *et al*, 2015); however, the enzyme responsible for this decoration is currently unknown.

In liverworts and mosses, side chains may additionally have branches of β1,4-linked galactose; in some higher plants backbone glucosyl residues have also been found to possess extra substitutions at the C2 hydroxyl (Pauly & Keegstra, 2016; Zavyalov *et al*, 2019). Decorations are not limited to glycosylation: acetylation can occur to unsubstituted backbone glucosyl residues (particularly in grasses) and on Gal/Ara*f* residues (Pauly & Keegstra, 2016). Finally, once in the wall, XyG may be hydrolysed and covalently cross-linked to other cell wall polysaccharides by xyloglucan endo-transglucosylases (XETs) (Franková & Fry, 2013). It is possible that XyG may also interact non-covalently with cellulose microfibrils (Park & Cosgrove, 2014).

1.6 Evolution, Taxonomy, and Phylogenetics

In 1973, the evolutionary biologist Theodosius Dobzhansky published an essay entitled *Nothing in Biology Makes Sense Except in the Light of Evolution* (Dobzhansky, 1973). In it, he states:

> *"Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a*

pile of sundry facts—some of them interesting or curious but making no meaningful picture as a whole."

Whilst his argument is actually concerned with the failure of 'antievolutionist' creationism to explain various natural phenomena, both title and quotation are just as well equipped to describe how we ought to interpret biochemical systems. Indeed, biochemical questions are answered more easily with an awareness of their evolutionary context. Therefore, this thesis will make frequent reference to the evolutionary relationships of glycosyltransferases and their host organisms. Hence, a brief summary of eukaryotic evolutionary relationships follows.

Eukaryota, the only domain of life to contain intracellular organelles, probably emerged roughly around two thousand million years ago (Eme *et al*, 2014). Sharing similar aspects with both Bacteria and Archæa, it is not exactly clear how the first eukaryotic common ancestor (FECA) came to be; nevertheless, by the time that the last eukaryotic common ancestor (LECA) came to exist, this organism possessed a nucleus, mitochondrion, endoplasmic reticulum, and, most relevantly, a Golgi apparatus (Klute *et al*, 2011; Bard, 2017).

Since that point in time, the eukaryotes have diverged into a number of different groups whose exact relationships are still being actively researched (Burki *et al*, 2020). Traditionally, the classification of living organisms (taxonomy) and the study of their ancestry (phylogenetics) have relied on biological and morphological characteristics; however, the modern age has seen a shift to computational techniques that analyse nuclear and organellar DNA sequences, and many of the most basal phylogenies are based on ribosomal genes (Stace, 1992; Keeling & Burki, 2019).

1.6.1 Animal evolution

Although the deepest roots of the eukaryotic tree are regularly revised, animals and their relatives are consistently placed with fungi (as well as some related amoebæ and slime moulds) in a group called Opisthokonta, which itself is currently considered to be a group of the Obazoa clade—in turn, part of the supergroup Amorphea (Adl *et al*, 2019; Burki *et al*, 2020). The opisthokonts can be divided into two groups: the Holomycota (or Nucletmycea), which include fungi, and the Holozoa, which include not only the multicellular Metazoa (*i.e.* animals), but also a grade of unicellular organisms descended from their ancestors—the most closely related

of these being the choanoflagellates (**Figure 1.14**) (Schalchian-Tabrizi *et al*, 2008; Adl *et al*, 2019).

The evolutionary transition from choanoflagellates to basal animals such as sponges and comb jellies represents an important lesson in how unicellular (albeit sometimes colony-forming) organisms can evolve to become obligately multicellular. Granted, multicellularity appears to have evolved independently in the eukaryotes multiple, perhaps dozens, of times (Grosberg & Strathmann, 2007); nevertheless, this particular transition can help us understand the underlying principles of animal development. Furthermore, the way in which animal multicellularity has been achieved could be rare amongst eukaryotes. For a start, as Cavalier-Smith (2017) explains, there are fundamentally two types mechanism: for walled cells, cell wall production can be modified such that cell division does not bring about full separation of daughter cells, or, for 'naked' cells such as those of animals and slime moulds, the cells can acquire a means to aggregate (note that extracellular glycans ought to be pivotal to both mechanisms). But, moreover, unlike autotrophs and saprotrophs, mobile phagotrophs (such as

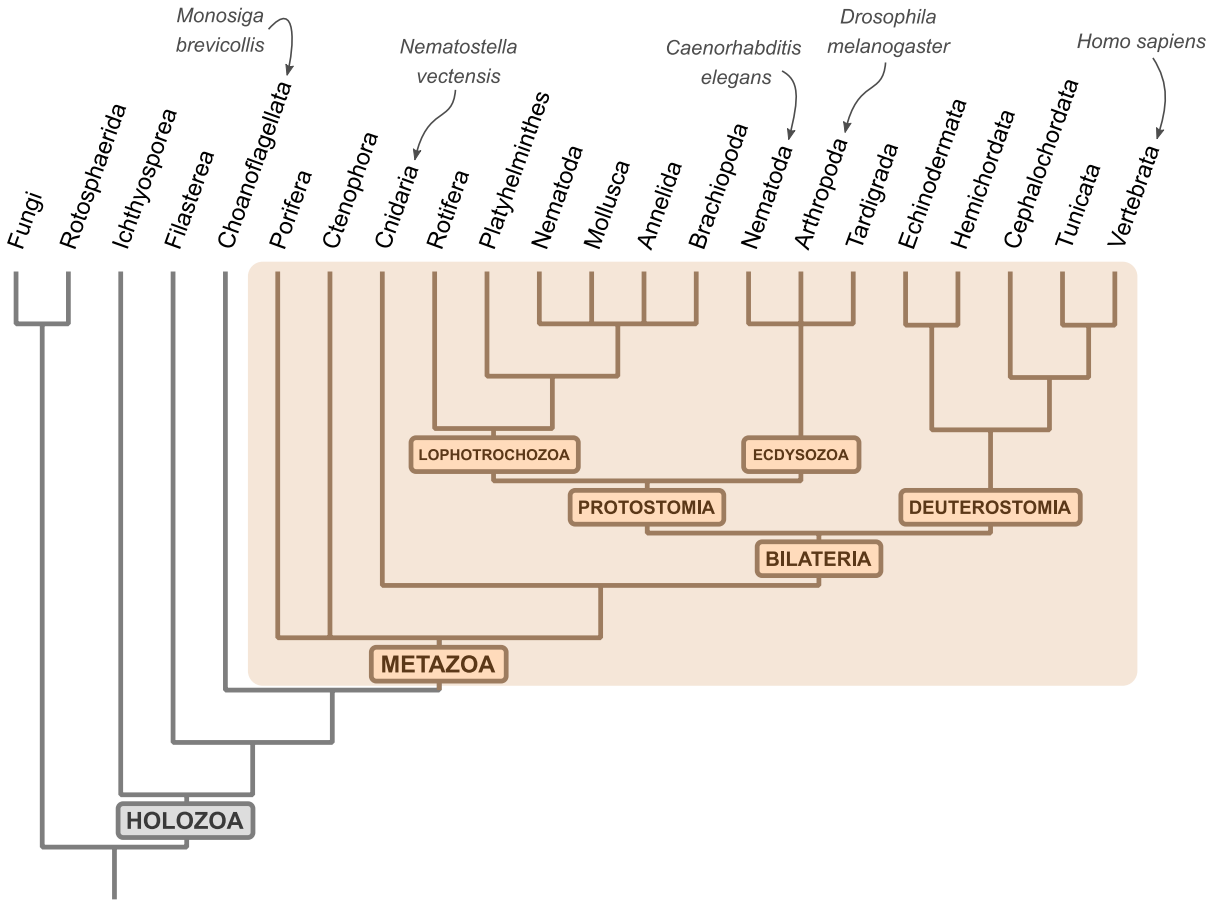


Figure 1.14 Simplified cladogram depicting evolutionary relationships between opisthokonts. Branch lengths are arbitrary. The brown box encompasses the 'animals'.

those constituted by animal ancestors) would ostensibly have faced a problem: most unicellular feeding strategies would be prohibited by obligate multicellularism. The answer to this, it would seem, lies in the shared feeding structures of choanoflagellates and sponges; both employ a 'collared' flagellum whose movements and structure draw in and capture nearby bacteria for phagocytosis (Cavalier-Smith, 2017; Brunet & King, 2017).

Despite this apparent similarity, the question of whether the most ancient metazoan lineage is constituted by Porifera (the sponges) or Ctenophora (the comb jellies) remains unresolved, though general morphological considerations seem to favour the former (Neff, 2018; Nielsen, 2019). Nevertheless, sponges have been studied as a simple model for cell–cell recognition and adhesion for well over a hundred years (Wilson, 1907; Misevic & Burger, 1993). At present, it is thought that sulphated proteoglycans play an important role in sponge cell adhesion; however, their glycan structures do not resemble those in other eukaryotes, and it seems that sulphated glycosaminoglycans are yet to be identified in any Porifera or Ctenophora species (despite the presence of exostosin-related genes in choanoflagellate genomes) (Yamada *et al*, 2011; Ori *et al*, 2011; Vilanova *et al*, 2016).

However, both HS and CS have been found in *Nematostella vectensis* and *Hydra magnipapillata*, which belong to Cnidaria—the next group to diverge from the metazoan tree (Feta *et al*, 2009; Yamada *et al*, 2011). These cnidarians are sister to a much larger group, the Bilateria, which in turn is split into Protostomia (which comprises many familiar invertebrates such as flatworms, molluscs, nematodes, and arthropods) and Deuterostomia (which comprises organisms such as starfish, sea cucumbers, lancelets, tunicates, and vertebrates) (Telford *et al*, 2015; Laumer *et al*, 2019). It has been suggested that deuterostomes exhibit particularly diversified glycan structures, with sialic acids and hyaluronan seldom found outside the group (Varki, 2011). However, in terms of other anionic polysaccharides, rather than being limited to deuterostomes, structural complexity may be a characteristic of marine animals by and large: molluscs can have particularly complex GAG structures for instance, and GAG sulphation in general appears to increase in organisms living in more saline environments (Yamada *et al*, 2011). Nevertheless, polysaccharides such as sulphated fucans and fucosylated chondroitin sulphate appear to be unique to echinoderms, while sulphated L-galactan may be specific to tunicates (Vasconcelos & Pomin, 2017; Mourão *et al*, 2018), though these species are all sea-dwelling. Furthermore, it appears that echinoderm *N*-glycans can also be highly sulphated (Vanbeselaere *et al*, 2020).

1.6.2 Plant evolution

In the broadest sense, plants constitute the Archæplastida—a completely separate eukaryotic supergroup from the Amorphea (Burki *et al*, 2020). The Archæplastida are usually considered to contain the Glaucophyta (blue-green algæ), the Rhodophyta (red algæ), and the Viridiplantæ (green algæ and plants); these groups have been traditionally thought to share a common ancestor—an early eukaryote that obtained the first plastid (a photosynthetic organelle derived from cyanobacterial endosymbiosis) (Keeling, 2013; Jackson *et al*, 2015). However, the robustness of the Archæplastida as a monophyletic group (*i.e.* one that contains all descendants of a single ancestor) remains to be confirmed by molecular phylogenetic techniques (Burki *et al*, 2020); furthermore, the evolutionary waters have been muddied by the discovery of secondary and tertiary endosymbiotic events that have resulted in the assimilation of plastids into completely unrelated taxa (brown algæ represent an example of such) (Keeling, 2013). Nonetheless, within the Viridiplantæ, recent genome and large-scale transcriptome sequencing projects now strongly support the existence of two clades: the Chlorophyta—which include many green algæ such as *C. reinhardtii* and *Volvox carteri*—and the Streptophyta, which comprises streptophyte algæ and land plants (Embryophyta; see **Figure 1.15**) (Leebens-Mack *et al*, 2019; Liang *et al*, 2020; Wang *et al*, 2020b).

The emergence of streptophyte algæ as a separate group probably coincided with a change to a freshwater habitat that was followed by the loss of sulphated extracellular glycans (which later marine plants have conspicuously since regained) (Aquino *et al*, 2011; Popper *et al*, 2011). The eventual cell wall compositions that emerged in these algæ vary considerably. The earliest-diverging branch, for instance, is currently thought to comprise the Mesostigmatophyceæ, the Chlorokybophyceæ, and the genus *Spirotænia* (Leebens-Mack *et al*, 2019; Wang *et al*, 2020b). *Chlorokybus atmophyticus*—a member of the Chlorokybophyceæ—possesses a cell wall that contains β1,3-glucan and a minor component of cellulose (as well as other, perhaps complex polysaccharides) (Sørensen *et al*, 2011). In contrast, instead of a polysaccharide-based wall, *Mesostigma viride* possesses proteinaceous scales with a carbohydrate constituent containing glucose and 3-deoxy-D-*lyxo*-heptulosaric acid (DHA) (Domozych *et al*, 1991). The next group to branch off the tree comprises the Klebsormidiophyceæ (Hori *et al*, 2014; Leebens-Mack *et al*, 2019; Cheng *et al*, 2019). The cell wall of *Klebsormidium nitens* appears be similar with that of *Chlorokybus atmophyticus*, except that it also appears to contain an arabinosylated β-1,4-xylan; supporting this, a *K. nitens* IRX10 homologue has been purified and shown to possess xylan β1,4-xylosyltransferase activity (Sørensen *et al*, 2011; Jensen *et al*, 2018; Hsieh

& Harris, 2019). The next three branches comprise the Charophyceæ, Coleochætophyceæ, and Zygnematophyceæ (Leebens-Mack *et al*, 2019; Cheng *et al*, 2019). Algæ from all three groups are thought to exhibit an overall cell wall composition similar to that of land plants, with cellulose, mixed linkage glucan, β1,4-mannan, homogalacturonan, β1,4-xylan, and xyloglucan, as well as potentially arabinogalactan, extensin, and rhamnogalacturonan I (Sørensen *et al*, 2010, 2011). However, the presence of homologues of various XyG-synthesising enzymes (such as CSLC, XXT, MUR3, FUT1, and AXY8) in the genome of *K. nitens* and the existence of XET activity in *K. crenulatum* suggests that xyloglucan may have emerged at least one step earlier in evolution (Herburger *et al*, 2018; Del-Bem, 2018).

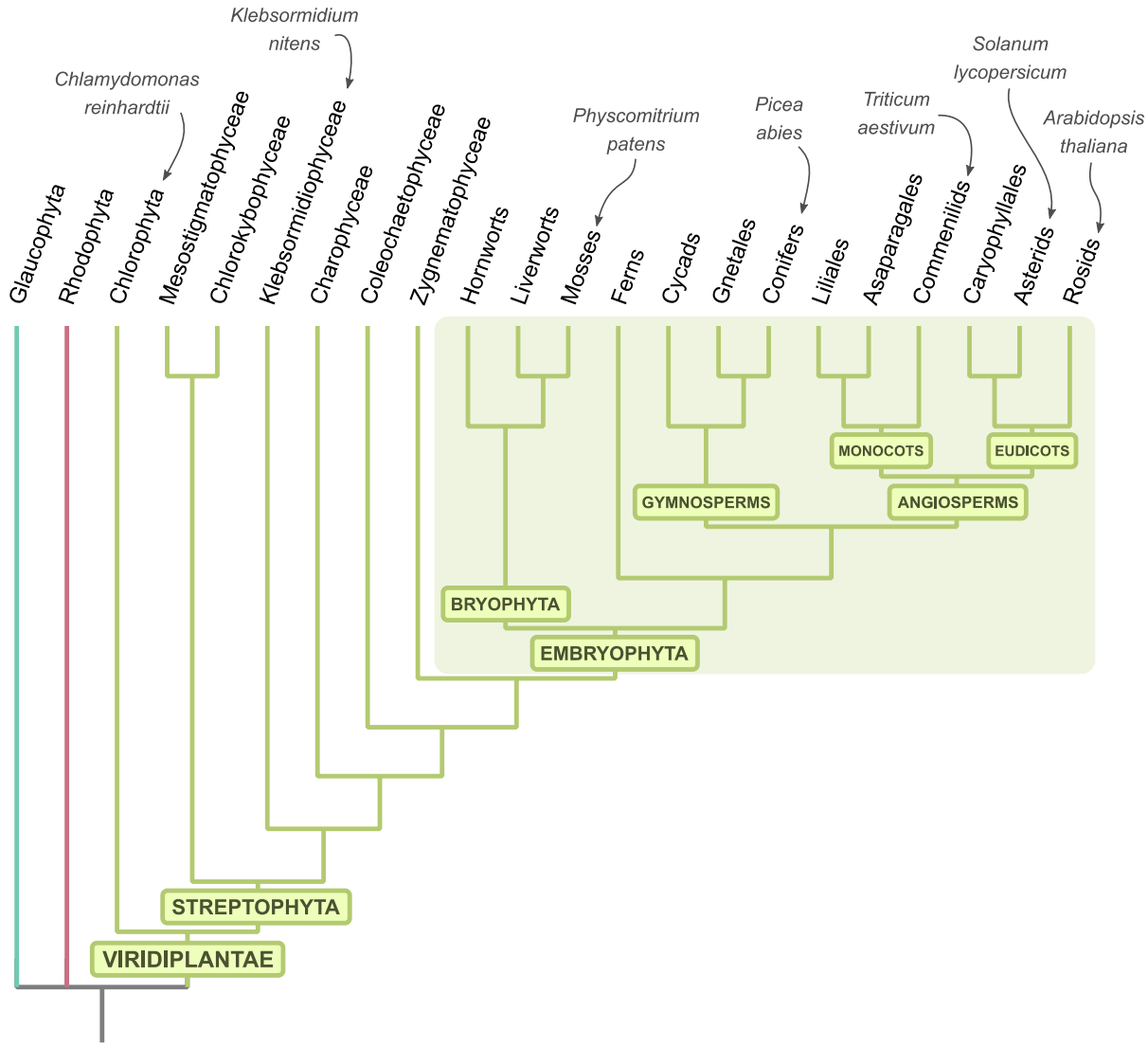


Figure 1.15 Simplified cladogram depicting evolutionary relationships between archæplastids. Branch lengths are arbitrary. The green box encompasses the 'land plants'.

The Zygnematophyceæ are considered the sister group of 'land plants' (though it is possible that, in reality, ancestral streptophyte algæ were already living terrestrially by the time these groups diverged) (Harholt *et al*, 2016; Del-Bem, 2018; Leebens-Mack *et al*, 2019; Cheng *et al*, 2019). The earliest-diverging land plants are the Bryophyta—currently thought to constitute a monophyletic group encompassing mosses (such as *P. patens*), liverworts (such as *Marchantia polymorpha*), and hornworts (such as *Anthoceros* spp.) (Puttick *et al*, 2018; Leebens-Mack *et al*, 2019; Zhang *et al*, 2020a; Li *et al*, 2020). Rhamnogalacturonan II appears to have emerged at the same time as the emergence of this group (Mikkelsen *et al*, 2014). Although these plants are not considered 'vascular plants', it has been recently demonstrated (for the first time beyond doubt) that mosses are able to transport water under tension using vascular structures (Brodribb *et al*, 2020). The 'traditional' vascular plants, designated 'Tracheophyta', have historically been distinguished by the presence of lignin in their vasculature (an innovation that might have supported vessel structure), though lignin production has arisen multiple times in non-vascular Archæplastida, and lignin-like species have now also been identified in bryophytes (Weng & Chapple, 2010; Popper *et al*, 2011; Novo-Uzal *et al*, 2012). The earliest branching tracheophytes are the Lycopodiopsida (lycophytes; such as the model organism *Selaginella moellendorffii*) followed by the Polypodiopsida (*i.e.* ferns) (Qi *et al*, 2018; Leebens-Mack *et al*, 2019). Sister to the Polypodiopsida are the Spermatophyta, or 'seed plants'. In turn, this group is divided into the gymnosperms (which include conifers) and the angiosperms (the flowering plants) (Magallon & Hilu, 2009). The angiosperms comprise an enormous number of species and, barring a small number of early-diverging species (*Amborella trichopoda*, for example, which is the most basal extant angiosperm), are further divided into 'monocots' and 'eudicots' (corresponding to the number of seedling cotyledons) (Soltis *et al*, 2019). Cell wall compositions vary substantially between the gymnosperms, monocots, and eudicots (as well as from tissue to tissue), though cellulose is always a major component (Pauly & Keegstra, 2008; Scheller & Ulvskov, 2010).

1.7 Rationale and aims

The ubiquity of complex extracellular glycans in living organisms is testament to their fundamental role in cell biology (Varki, 2011). Moreover, glycan function is an important subject of study because human and pathogen carbohydrates play a central role in the course of cancer and infectious diseases (Schnaar, 2016; Reily *et al*, 2019). Furthermore, although plant carbohydrates are an essential source of nutrition, we are still in the process of learning about the complex interactions between dietary carbohydrates and symbiotic gut microbiota

(Ndeh & Gilbert, 2018; Zmora *et al*, 2019); hence, more research is required into the health properties and structural diversity of plant polysaccharides. Likewise, we rely on plant fibres for materials and fuel, yet biochemical discoveries could produce innovative science-based applications (Pauly & Keegstra, 2008; Sorieul *et al*, 2016).

Glycan structures are largely determined by the activity of glycosyltransferases, and defects in GT-encoding genes cause congenital disorders of glycosylation (CDG) in humans (Jaeken & Péanne, 2017). The characterisation of GT activities and structures not only helps to explain disease mechanisms but can also permit the prediction, engineering, and synthesis of undiscovered and novel glycans with therapeutic or useful material properties. In particular, the GT47 family is a worthy target for study because, in addition to its importance for human development and health, it comprises a large number of enzyme activities involved in plant cell wall synthesis (Geshi *et al*, 2011; Busse-Wicher *et al*, 2014). Despite this, many plant GT47s are yet to be functionally characterised, and at present there is no published high-resolution structural data of any GT47 to aid prediction of substrate specificity or reaction mechanism; nor is there enough structural information to establish a detailed disease mechanism in the human GT47s (Awad *et al*, 2018). In addition, the synthesis of xylan—one of the most abundant polysaccharides on the planet (Deutschmann & Dekker, 2012)—remains a mysterious process due to our ignorance of certain GT47 and GT43 functions, impeding efforts to engineer plant cell walls.

Glycosyltransferases are the bread and butter of the Golgi proteome. Despite the fact that the Golgi apparatus is one of scarce features uniting virtually all eukaryotic cells, this organelle remains rather opaque with regard to many aspects of cell biology (Klute *et al*, 2011; Bard, 2017; Pothukuchi *et al*, 2019; McCaughey & Stephens, 2019). For instance, our understanding of Golgi targeting mechanisms has historically lagged behind that of other cell compartments, while the numerous reports of GT oligomerisation and sub-Golgi localisation hint at a nano-scale organisation that has not yet been thoroughly investigated (Schoberer & Strasser, 2011; Kellokumpu *et al*, 2016; Welch & Munro, 2019; Parsons *et al*, 2019; McCaughey & Stephens, 2019). Therefore, the study of GT localisation and oligomerisation mechanisms (as well as the functioning of multi-domain GTs) will advance our understanding of this fascinating eukaryotic structure.

The scientific aims comprised in each results chapter are given below.

Transmembrane dimerisation of IRX9 and IRX14 (Chapter 3)

In preliminary work, conserved sequence motifs were identified in the predicted transmembrane helix regions of GT43 members IRX9 and IRX14 that could potentially facilitate self-interaction; aside from disulphide-forming cysteine residues, a specific motif-driven transmembrane interaction has never been reported for a GT before. The main aim of this project was to evaluate the potential role of various residues in these proteins in GT oligomerisation by both bioinformatic and biochemical methods. In short, the full aims were to:

- Analyse the conservation of residues in both IRX9 and IRX14, as well as the potential of these residues to contribute to homodimerisation, by bioinformatic methods
- Confirm the self-interaction of IRX9 and IRX14 TM peptides heterologously expressed in *E. coli* using the TOXGREEN reporter assay
- Analyse the ability of full-length IRX9 TM point mutants to function *in planta* in place of the wild-type protein
- Determine the effect of TM point mutation on localisation and expression of IRX9 by transient expression of GFP-fusions in tobacco leaves

Structure and activity of exostosin-like 3 (Chapter 4)

Many lines of inquiry into the functions of GT47 glycosyltransferases are impeded by the lack of protein structural data. In addition, the exact purpose of evolutionary GT–GT fusion events is unknown since no such eukaryotic protein has been structurally characterised. However, purification and crystallisation of eukaryotic GTs is particularly difficult and often fails. The GT47–GT64 fusion constituted by EXTL3 poses an ideal target for investigation because it is just large enough to attempt cryo-electron microscopy (cryo-EM). However, the GT47 domain of EXTL3 is reported to be inactive. Therefore, the aims of this project were to verify the activities of EXTL3 and to solve the atomic structure of EXTL3. In short:

- Develop an assay for heparan backbone extension activity
- Determine the activities of EXTL3
- Solve the structure of EXTL3 by cryo-EM (with collaboration)
- Solve the cryo-EM structure of EXTL3 bound to various substrates
- Investigate the evolution of GT47 and GT64 activities using bioinformatic methods

Nucleotide sugar specificity of xylan glucuronic acid pyranosyltransferases (Chapter 5)

The atomic structure of EXTL3 presented the opportunity to investigate the structural determinants of substrate specificity in plant GT47-A enzymes. For example, the recently discovered *Eg*XAPT1 and *Eg*XLPT1 enzymes are highly similar in sequence and closely related in activity yet differ in donor specificity. The aim of this project was to explain the difference in substrate specificity using the amino acid sequences and structural models of these enzymes. In short:

- Sort plant GT47-A sequences into robust clades in order to identify further XAPT orthologues
- Compare XAPT homologue sequences with a structural model to propose particular amino acids important for donor specificity
- Determine when the ancestral XAPT was duplicated to give rise to XAPT and XLPT in *Eucalyptus spp.* (employing phylogenetic analysis of not only published sequences but also those amplified by PCR from genomic DNA extracted from relevant species)
- Analyse the xylan structures of relevant species to look for a correlation between XAPT/XLPT sequence elements and arabinopyranosyl *vs* galactosyl decorations
- Test a leading hypothesis by determining the activity of a *Eg*XAPT1 point mutant expressed in Arabidopsis

Nucleotide sugar specificity in the wider GT47-A clade (Chapter 6)

Hypotheses about XAPT substrate specificity might also be extended to the rest of the GT47-A clade. The aims of the project were to analyse the possible determinants of donor specificity across the whole group and to predict and verify the existence of novel enzyme activities within it. In short:

- Investigate the correlation between the sequence of the Nβ5–Nα5 loop and substrate specificity in GT47-A enzymes
- Identify candidates that might exhibit a novel activity based on their Nβ5–Nα5 loop sequences
- Determine the ability of candidate enzymes to complement the phenotypes of Arabidopsis GT47 xyloglucan mutants
- Develop methods for structural analysis of xyloglucan
- Determine the activity of novel xyloglucan-specific GT47 enzymes

Chapter 2 : Materials and methods

2.1 Bioinformatics

2.1.1 Alignments and phylogenies of GT43 sequences

GT43 sequences were downloaded from the PlantCAZyme database (http://bcb.unl.edu/plantcazyme/) (Ekstrom *et al*, 2014). Redundant isoforms were removed manually. After combination with previously characterised GT43 sequences from *Klebsormidium nitens* (Taujale & Yin, 2015; Jensen *et al*, 2018), these sequence were aligned with MUSCLE (Edgar, 2004a, 2004b) before creation of a hidden Markov model (HMM) using HMMER (Eddy, 2008, 2009, 2011). HMM searches to *Mesotænium endlicherianum* (Cheng *et al*, 2019), *Anthoceros angustus* (Zhang *et al*, 2020a), and *Salvinia cucullata* (Li *et al*, 2018) proteome models were performed using hmmsearch with an $E \leq 10^{-40}$ cut-off. After the addition of the GlcAT-I protein sequence, the sequences were aligned with MAFFT (Katoh & Standley, 2013), and a phylogeny was constructed using FastTreeMP (Price *et al*, 2010). The GT43-A and GT43-B subtrees were extracted using FigTree (Rambaut & Drummond, 2018); sequences were extracted from the original FASTA file by virtue of their IDs using a custom awk script (see **Appendix**). Two *Zea mays* sequences and one *Gossypium raimondii* sequence were removed due to their short length. A custom Python script (`TMHgrab.py`; see **Appendix**), which invokes TMHMM v1 (Krogh *et al*, 2001) via the command-line, was used to extract TMH regions from these sequences (IRX9 from *K. nitens* was truncated to permit appropriate TMH detection). These TMH sequences were then aligned using MUSCLE, with a gap opening penalty of −30 for GT43-A and −15 for GT43-B sequences to prevent the formation of gaps within the transmembrane region.

2.1.2 Alignments of EXTL3 orthologues

To obtain EXTL3 orthologues, metazoan and choanoflagellate reference proteomes were searched using the online phmmer server (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer) (Potter *et al*, 2018) with human EXTL3 as the input and an $E \leq 10^{-130}$ cut-off (in order to exclude closely related EXT1 and EXT2 sequences). The identified sequences were aligned using MAFFT, with subsequent removal of sequences with insertions or deletion of more than five residues in the C-terminal region.

2.1.3 Phylogeny of GT47-family sequences from animals, M. brevicollis*, and plants*

The GT47 HMM was downloaded from the dbCAN2 database (http://bcb.unl.edu/dbCAN2/) (Yin *et al*, 2012; Zhang *et al*, 2018a). From NCBI Genome (https://www.ncbi.nlm.nih.gov/genome/), proteome models for *Homo sapiens*, *Drosophila melanogaster*, and *Cænorhabditis elegans* were downloaded. In addition, proteome models for *Arabidopsis thaliana* and *Physcomitrium patens* from JGI Phytozome v12 (https://phytozome.jgi.doe.gov/pz/portal.html), *Amphimedon queenslandica* from EnsemblMetazoa (https://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index), *Monosiga brevicollis* MX1 (King *et al*, 2008) from JGI MycoCosm (https://mycocosm.jgi.doe.gov/Monbr1/Monbr1.home.html), and *Ginkgo biloba* (Guan *et al*, 2016, 2019) from CNGBdb (http://gigadb.org/dataset/100613) were downloaded. These proteome models were searched using the GT47 HMM with a cut-off of $E \leq 10^{-10}$. Redundant isoforms were removed manually. Furthermore, GT64 and sulphatase domains were removed with reference to their PFAM predictions. All sequences were then aligned using MAFFT. Pp3c25_4460V3.1 was removed due to poor alignment. Each sequence was truncated to its GT47 domain (corresponding to residues 196–538 of human EXTL3) using a custom Python script (`DomainExtract.py`, **Appendix**). ProtTest3 (Darriba *et al*, 2011) was then used to determine an appropriate substitution model (WAG+Γ+I+F) before construction of a tree with RAxML (Stamatakis, 2014) with 100 rapid bootstrap replicates.

2.1.4 Phylogeny of GT64-family sequences from animals, M. brevicollis*, and plants*

GT64 sequences were obtained in an identical manner to GT47 sequences, substituting a dbCAN2 GT64 HMM for the GT47 HMM. Sequences were aligned with MUSCLE. Aqu2.1.11978_001, Aqu2.1.36600_001, and Aqu2.1.36602_001 were removed from the alignment due to their short length. Each sequence was truncated to its GT47 domain (corresponding to residues 663–919 of human EXTL3) using `DomainExtract.py`. ProtTest3 determined LG+Γ+I to be a suitable substitution model; subsequently, a phylogeny was constructed with RAxML with 100 rapid bootstrap replicates.

2.1.5 Phylogeny of the GT47-A clade

All sequences from the MUR3 homologue clusters curated by the plant comparative genomic platforms PLAZA Dicots 4.5, PLAZA Monocots 4.5, and PLAZA Gymnosperms 3.0 (Proost *et al*, 2015; Van Bel *et al*, 2018) were downloaded and combined. The 39 GT47-A sequences from Arabidopsis were then aligned using MUSCLE and used to create an HMM using HMMER. A combination of hmmsearch (using the *At*GT47-A HMM) and TBLASTN

(Altschul *et al*, 1990, 1997) searches (using *At*MUR3 as bait) was then used to obtain additional GT47-A sequences from the proteome models/untranslated genomes of *Klebsormidium nitens* (Hori *et al*, 2014), *Sphagnum fallax* (Weston *et al*, 2018), *Azolla filiculoides* and *Salvinia cucullata* (Li *et al*, 2018), *Nymphæa colorata* (Zhang *et al*, 2020b), *Liriodendron chinense* (Chen *et al*, 2019), and *Aquilegia coerulea* 'Goldsmith' (Filiault *et al*, 2018) (see **Table 2.1** for the final list of species). All sequences were then aligned using MAFFT before truncation to their GT47 domain using `DomainExtract.py`. Poorly aligned sequences were removed by hand. ProtTest3 was then used to establish a suitable evolution model (LG). Subsequently, the phylogeny was constructed using FastTreeMP (Price *et al*, 2010) with 100 bootstrap replicates. Pseudo-replicates were created using SEQBOOT from the PHYLIP package (Felsenstein, 1989). The phylogeny was visualised using FigTree (Rambaut & Drummond, 2018).

2.1.6 Phylogeny of the XAPT subclade

The aligned sequences of XAPT homologues were extracted from the relevant subtree of the larger GT47-A tree using FigTree and awk. A XAPT HMM was then constructed using HMMER. HMMER hmmsearch (using the XAPT HMM) was used to obtain additional XAPT homologues from the proteome models of *Metrosideros polymorpha* (Izuno *et al*, 2016), *Syzygium oleosum*, *Rhodamnia argantea*, and *Punica granatum* (Yuan *et al*, 2018; Luo *et al*, 2020), all available from NCBI Genome. TBLASTN was also used to search for homologues from a *Oenothera rosea* transcriptome (Leebens-Mack *et al*, 2019). Sequences were aligned using MUSCLE before truncation to the GT47 domain. ProtTest3 was then used to establish a suitable evolution model (JTT+Γ+I). A phylogeny was then constructed using RAxML with 100 rapid bootstraps. The phylogeny was visualised using FigTree.

2.1.7 Modelling of At*XAPT1 structure*

The sequence of *At*XAPT1 was truncated to its predicted GT47 domain before submission to the I-TASSER server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) (Zhang *et al*, 2008; Roy *et al*, 2010; Yang *et al*, 2015) with the GT47 domain of the apo-EXTL3 structure provided as a template. The highest-scoring model was aligned to the EXTL3 structure (which was itself aligned to other GT-B structures using DALI) using PyMOL. Alignments of the *At*XAPT, *Eg*XAPT, and *Eg*XLPT protein sequences were made using MUSCLE, and residue differences were manually annotated on the *At*XAPT1 model structure.

Table 2.1 Species involved in the creation of the GT47-A tree. GT47-A sequences were obtained from a genome, transcriptome, or proteome model of each species.

Species	Source
Streptophyte algæ grade	
Klebsormidium nitens	Klebsormidium Genome Project (Hori *et al*, 2014); http://www.plantmorphogenesis.bio.titech.ac.jp/~algae_genome_project/klebsormidium/
Bryophyta	
Physcomitrium patens	Dicots PLAZA 4.5 (Van Bel *et al*, 2018); https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v4_5_dicots/
Marchantia polymorpha	Dicots PLAZA 4.5
Sphagnum fallax	*Sphagnum fallax v0.5* at JGI Phytozome v12 (Weston *et al*, 2018); https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sfallax
Lycopodiophyta	
Selaginella moellendorffii	Dicots PLAZA 4.5
Polypodiopsida	
Azolla filiculoides	NCBI Genome; Li *et al* (2018)
Salvinia cucullata	NCBI Genome; Li *et al* (2018)
Gymnospermæ	
Ginkgo biloba	Gymno PLAZA 1.0 (Proost *et al*, 2015); https://bioinformatics.psb.ugent.be/plaza/versions/gymno-plaza/
Cycas micholitzii	Gymno PLAZA 1.0
Taxus baccata	Gymno PLAZA 1.0
Gnetum montanum	Gymno PLAZA 1.0
Pseudotsuga menziesii	Gymno PLAZA 1.0
Pinus tæda	Gymno PLAZA 1.0
Pinus sylvrestris	Gymno PLAZA 1.0
Pinus pinaster	Gymno PLAZA 1.0
Picea sitchensis	Gymno PLAZA 1.0
Picea abies	Gymno PLAZA 1.0
Picea glauca	Gymno PLAZA 1.0
Angiospermæ	
—ANA grade	
Amborella trichopoda	Dicots PLAZA 4.5
Nymphæa colorata	NCBI Genome; Zhang *et al* (2020b)
Liriodendron chinense	NCBI Genome; Chen *et al* (2019)
—Monocotidæ	

Zostera marina	Monocots PLAZA 4.5 (Van Bel *et al*, 2018)
Spirodela polyrhiza	Monocots PLAZA 4.5
Asparagus officinalis	Monocots PLAZA 4.5
Apostasia shenzhenica	Monocots PLAZA 4.5
Phalænopsis equestris	Monocots PLAZA 4.5
Elæis guineensis	Monocots PLAZA 4.5
Musa acuminata	Monocots PLAZA 4.5
Calamus simplicifolius	Monocots PLAZA 4.5
Ananas comosus	Monocots PLAZA 4.5
——Poaceæ	
Zea mays B104	Monocots PLAZA 4.5
Zea mays PH207	Monocots PLAZA 4.5
Zea mays B73	Monocots PLAZA 4.5
Miscanthus sinensis	Monocots PLAZA 4.5
Saccharum spontaneum	Monocots PLAZA 4.5
Sorghum bicolor	Monocots PLAZA 4.5
Cenchrus americanus	Monocots PLAZA 4.5
Setaria italica	Monocots PLAZA 4.5
Zoysia japonica spp. nagirizaki	Monocots PLAZA 4.5
Oropetium thomæum	Monocots PLAZA 4.5
Oryza sativa ssp. indica	Monocots PLAZA 4.5
Oryza sativa ssp. japonica	Monocots PLAZA 4.5
Oryza brachyantha	Monocots PLAZA 4.5
Phyllostachys edulis	Monocots PLAZA 4.5
Lolium perenne	Monocots PLAZA 4.5
Triticum æstivum	Monocots PLAZA 4.5
Triticum turgidum	Monocots PLAZA 4.5
Hordeum vulgare	Monocots PLAZA 4.5
Brachypodium distachyon	Monocots PLAZA 4.5
—Eudicotidæ	
——Basal eudicots	
Aquilegia coerulia	NCBI Genome; Filiaut *et al* (2018)
Nelumbo nucifera	Dicots PLAZA 4.5
——Superrosidæ	
Vitis vinifera	Dicots PLAZA 4.5
Populus trichocarpa	Dicots PLAZA 4.5
Ricinus communis	Dicots PLAZA 4.5
Manihot esculenta	Dicots PLAZA 4.5
Hevea brasiliensis	Dicots PLAZA 4.5
Arachis ipænsis	Dicots PLAZA 4.5
Vigna radiata var. *radiata*	Dicots PLAZA 4.5

Glycine max	Dicots PLAZA 4.5
Cajanus cajan	Dicots PLAZA 4.5
Cicer arietinum	Dicots PLAZA 4.5
Trifolium pratense	Dicots PLAZA 4.5
Medicago trunculata	Dicots PLAZA 4.5
Citrullus lanatus	Dicots PLAZA 4.5
Cucumis sativus L.	Dicots PLAZA 4.5
Cucumis melo	Dicots PLAZA 4.5
Ziziphus jujuba	Dicots PLAZA 4.5
Fragaria vesca	Dicots PLAZA 4.5
Prunus persica	Dicots PLAZA 4.5
Pyrus bretschneideri	Dicots PLAZA 4.5
Malus domestica	Dicots PLAZA 4.5
Eucalyptus grandis	Dicots PLAZA 4.5
Citrus clementina	Dicots PLAZA 4.5
Gossypium raimondii	Dicots PLAZA 4.5
Theobroma cacao	Dicots PLAZA 4.5
Corchorus olitorius	Dicots PLAZA 4.5
Carica papaya	Dicots PLAZA 4.5
Tarenaya hassleriana	Dicots PLAZA 4.5
Schrenkiella parvula	Dicots PLAZA 4.5
Brassica oleracea	Dicots PLAZA 4.5
Brassica rapa	Dicots PLAZA 4.5
Capsella rubella	Dicots PLAZA 4.5
Arabidopsis thaliana	Dicots PLAZA 4.5
Arabidopsis lyrata	Dicots PLAZA 4.5
——Caryophyllales	
Beta vulgaris	Dicots PLAZA 4.5
Chenopodium quinoa	Dicots PLAZA 4.5
Amaranthus hypochondriacus	Dicots PLAZA 4.5
——Superasteridæ	
Actinidia chinensis	Dicots PLAZA 4.5
Daucus carota	Dicots PLAZA 4.5
Coffea canephora	Dicots PLAZA 4.5
Utricularia gibba	Dicots PLAZA 4.5
Erythranthe guttata	Dicots PLAZA 4.5
Petunia axillaris	Dicots PLAZA 4.5
Solanum lycopersicum	Dicots PLAZA 4.5
Solanum tuberosum	Dicots PLAZA 4.5
Capsicum annuum	Dicots PLAZA 4.5

2.1.8 Phylogeny of the XAPT and XLPT homologues in the Myrtaceæ

Supplementary genomic sequences of XAPT and XLPT were identified using hmmsearch and TBLASTN, as above, from the genomes of *Eucalyptus camaldulensis*, *E. melliodora*, *Corymbia citriodora*, *Angophora floribunda*, and *Psidium guajava*, available from NCBI Genome. Sequences were aligned using MACSE (Ranwez *et al*, 2011, 2018) and truncated to a region containing 102 codons using a custom Python script. A tree was then constructed, under a MGK+F1X4+G4 substitution model (as suggested by ModelFinder), using IQ-TREE v1.6 with 1000 ultrafast bootstraps (Nguyen *et al*, 2015; Kalyaanamoorthy *et al*, 2017; Hoang *et al*, 2018).

2.1.9 Nβ5–Nα5 loop alignment

The original MAFFT alignment of GT47-A sequences was truncated to the eight columns corresponding to the predicted Nβ5–Nα5 in *At*XAPT1 (Arg204–Arg211) using `DomainExtract.py`. Any sequences with indels in this region were removed from the alignment. The sequence identifiers from each cluster were extracted from the GT47-A tree using FigTree, and awk was used to extract the corresponding Nβ5–Nα5 loop sequences from the alignment.

2.1.10 Phylogeny of GT47-A sequences from Ericales genomes

Multiple TBLASTN searches, each using one of the 39 Arabidopsis GT47-A sequences as a query, were used to search for GT47-A sequences from the genomes of *Vaccinium macrocarpon* (Polashock *et al*, 2014), *V. corymbosum* (Bian *et al*, 2014), and *Argania spinosa*, available from NCBI Genome (*E* value threshold = 1×10^{-60}). Redundant hits were removed using CD-HIT with a 95 % identity threshold. I also used the *At*GT47-A HMM described above to search for GT47-A sequences from the proteome models of *Rhododendron williamsianum* (Soza *et al*, 2019) and *Camellia sinensis* (Wei *et al*, 2018). Sequences were then aligned using MUSCLE. After truncation to the GT47 domain, ProtTest3 was used to determine an appropriate substitution model (JTT+Γ+I). A phylogeny was then constructed using RAxML, with 100 rapid bootstrap replicates.

2.1.11 Sequence logos

All sequence logos were created using WebLogo 3 (http://weblogo.threeplusone.com) (Crooks *et al*, 2004), except in the case of the IRX9 TMH experiments, which used WebLogo version 2.8.2 (https://weblogo.berkeley.edu).

2.2 Plant genotypes, growth, crossing, and photography

All Arabidopsis lines were of the Columbia-0 ecotype. The *irx9-1* mutant (SALK_058238) was a gift of Jan Lyczakowski (University of Cambridge). The *mur3-1* (CS8566), *mur3-3* (SALK_141953), and *xlt2* (GABI_552C10) mutants were gifts of Li Yu (University of Cambridge). Before germination of transgenic lines, seeds were sterilised in 1 ml sterilisation solution comprising 80 % ethanol, 1 % household bleach, and 1 % Tween-20 for 10 min in a benchtop vortex shaking at 1200 rpm before successive washes with 65 %, 80 %, and 100 % ethanol and air drying under sterile conditions. These seeds were transferred onto MS plates (2.6 mM MES-KOH, *p*H 5.6–5.8, 0.22 % Murashige and Skoog Basal Salt Mixture, 1 % sucrose, and 0.8 % plant agar) and stratified in the dark at 4 °C for three days before transfer to a 16 h/8 h photoperiod at room temperature. After 7–14 days, seedlings were transferred to soil (a 9:1 mix of Levington M3 compost and vermiculite) and grown under a 16 h/8 h photoperiod at 21 °C with weekly watering. Non-transgenic Arabidopsis lines and tobacco (*Nicotiana benthamiana*) were sown directly to soil and grown under the same conditions. Arabidopsis plants were grown to an age of 6–7 weeks for photography before ceasing watering; plants were allowed to fully senesce before collection of seeds and material for the creation of AIR (see below). Tobacco was grown for 3–4 weeks prior to infiltration.

Plant crosses to produce the *xlt2 mur3-1* double mutant were achieved as described by Weigel & Glazebrook (2006) with assistance from Xiaolan Yu. Verification of the genotype was achieved as previously described (Jensen *et al*, 2012), except that the *Hinf*I restriction enzyme was used in the dCAPS method (Neff *et al*, 2002) (primers were designed accordingly).

Plants in the same experiment were photographed on the same day. Images were cropped and the contrast improved using ImageJ (Schneider *et al*, 2012).

2.3 Molecular biology

2.3.1 Golden Gate MoClo assembly

Expression vectors for Arabidopsis and tobacco were created using Golden Gate MoClo technology as described by Weber *et al* (2011) and Patron *et al* (2015), with thermocycler reactions adapted from Engler *et al* (2009) (**Table 2.2 and Table 2.3**). Plasmid concentrations were determined using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher). A visual summary of the general Golden Gate cloning strategy is shown in **Figure 2.1**. Relevant Golden Gate modules and assemblies are listed in **Tables 2.4–2.6**. Level 0 coding sequences were synthesised commercially—except in the case of the *Sl*XST1 coding sequence, which was

amplified from *Solanum lycopersicum* genomic DNA by PCR with Q5® High-Fidelity DNA Polymerase (NEB) (forward primer: 5'-GTGGTCTCAAATGTTGCCATCTGAAAATTCTTCCCC-3'; reverse primer: 5'-GTGGTCTCACGAACTTAGTTTTTGTTGCTTGAATCTC-3'). Correct assembly was verified by sequencing by the DNA Sequencing Facility (Department of Biochemistry, University of Cambridge).

Table 2.2 MoClo assembly reagent quantities.

Volume	Reagent	Concentration
1 µl	vector backbone (in PCR-grade water)	100 ng μl^{-1}
n × 1 µl	each additional MoClo part (in dH_2O)	100 ng μl^{-1}
1.5 µl	T4 DNA ligase buffer (NEB)	(10×)
0.15 µl	BSA (for level 1 assembly only)	(100×)
1 µl	*Bsa*I-HF®v2 (for level 1) or *Bbs*I-HF® (for level 2) (NEB)	20 kU ml^{-1}
1 µl	T4 DNA ligase (NEB)	400 kU ml^{-1}
to 15 µl	PCR-grade water	

Table 2.3 MoClo assembly thermocycler conditions.

Step	Cycles
37 °C for 3 min 16 °C for 4 min	×25
50 °C for 5 min 80 °C for 5 min	×1

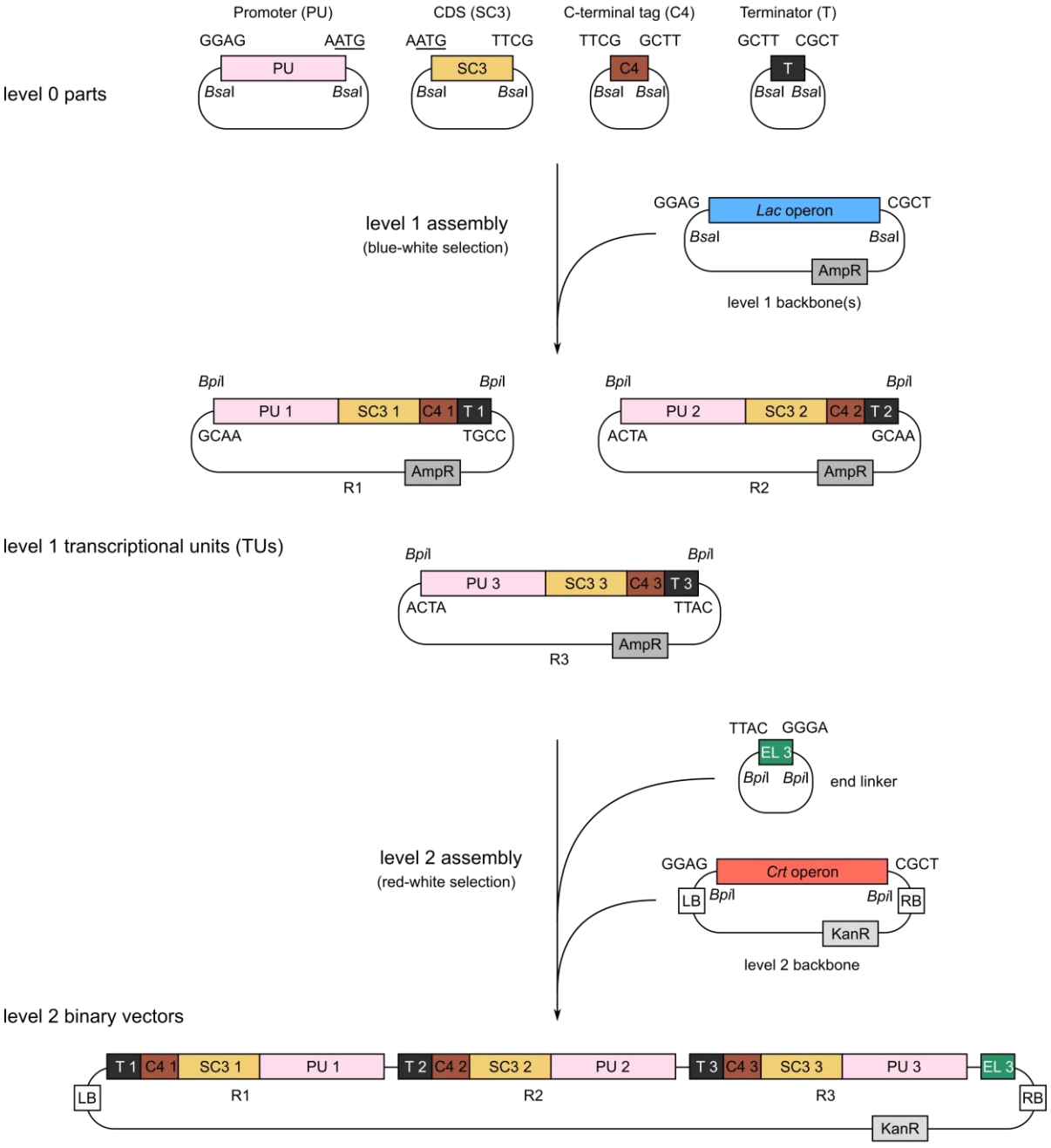


Figure 2.1 Dupree lab Golden Gate assembly flowchart.

Table 2.4 Pre-existing / level 0 MoClo parts used in assemblies.

Name	Part type	Part contents/function	Source
EC16527	promoter	4×p35S	Commercial MoClo kit (Engler *et al*, 2014)
EC62016	promoter	pIRX3	Lab part (PCR from Arabidopsis gDNA)
EC62031	promoter	pCESA3	Lab part (PCR from Arabidopsis gDNA)
EC62311	promoter	pXXT2	Lab part (PCR from Arabidopsis gDNA)
EC62312	promoter	pIRX14	Lab part (PCR from Arabidopsis gDNA)
EC62080	coding sequence	*Eg*XAPT1	Synthesised *de novo* by GeneWiz
EC62062	coding sequence	mannosidase-I N-terminus	Synthesised *de novo* by GeneWiz
EC62053	coding sequence	*At*IRX9	Synthesised *de novo* by GeneWiz
EC62055	coding sequence	*At*IRX10	Synthesised *de novo* by GeneWiz
EC62057	coding sequence	*At*IRX14	Synthesised *de novo* by GeneWiz
EC62326	coding sequence	FVE16326	Synthesised *de novo* by GeneWiz
EC62332	coding sequence	*At*GT19	Synthesised *de novo* by GeneWiz
EC62333	coding sequence	*At*XUT1	Synthesised *de novo* by GeneWiz
EC62334	coding sequence	Bv7_174350	Synthesised *de novo* by GeneWiz
EC62336	coding sequence	*Vm*GT47-A12	Synthesised *de novo* by GeneWiz
EC62337	coding sequence	Cc07_g06550	Synthesised *de novo* by GeneWiz
EC62337	coding sequence	Cc07_g06570	Synthesised *de novo* by GeneWiz
EC62303	coding sequence	*At*IRX9[G28I]	Obtained by site-directed mutagenesis of EC62053
EC62306	coding sequence	*At*IRX9[C24S]	Obtained by site-directed mutagenesis of EC62053
EC62341	coding sequence	*Eg*XAPT[A235G]	Obtained by site-directed mutagenesis of EC62080
EC41421	terminator	NosT	Commercial MoClo kit (Engler *et al*, 2014)
EC16565	C-terminal tag	mCherry	Commercial MoClo kit (Engler *et al*, 2014)
EC16570	C-terminal tag	eGFP	Commercial MoClo kit (Engler *et al*, 2014)
EC16572	C-terminal tag	3×HA	Commercial MoClo kit (Engler *et al*, 2014)
EC16573	C-terminal tag	FLAG	Commercial MoClo kit (Engler *et al*, 2014)
EC16574	C-terminal tag	3×Myc	Commercial MoClo kit (Engler *et al*, 2014)
EC47802	backbone	level 1; R1 position	Commercial MoClo kit (Engler *et al*, 2014)
EC47811	backbone	level 1; R2 position	Commercial MoClo kit (Engler *et al*, 2014)
EC47822	backbone	level 1; R3 position	Commercial MoClo kit (Engler *et al*, 2014)
EC47831	backbone	level 1; R4 position	Commercial MoClo kit (Engler *et al*, 2014)
EC50506	backbone	level 2	Commercial MoClo kit (Engler *et al*, 2014)
EC41722	level 2 end linker	end linker for 1 TU	Commercial MoClo kit (Engler *et al*, 2014)
EC41744	level 2 end linker	end linker for 2 TUs	Commercial MoClo kit (Engler *et al*, 2014)
EC41766	level 2 end linker	end linker for 3 TUs	Commercial MoClo kit (Engler *et al*, 2014)
EC41780	level 2 end linker	end linker for 4 TUs	Commercial MoClo kit (Engler *et al*, 2014)
EC62111	transcriptional unit	pOleosin::OLE1-eGFP::ActinT	Lab part (constructed from level 0 parts)
EC62112	transcriptional unit	4×p35S::KanR::35ST	Lab part (constructed from level 0 parts)

Table 2.5 Level 1 transcriptional units assembled in this work

Name		Backbone	Inserted parts			
EC62197	4×p35S::ManI-mCherry::NosT	EC47802 (R1)	EC16527	EC62062	EC16565	EC41421
EC62198	4×p35S::IRX10-FLAG::NosT	EC47811 (R2)	EC16527	EC62055	EC16573	EC41421
EC62199	4×p35S::IRX14-HA::NosT	EC47822 (R3)	EC16527	EC62057	EC16572	EC41421
EC62400	4×p35S::IRX9-GFP::NosT	EC47831 (R4)	EC16527	EC62053	EC16570	EC41421
EC62402	4×p35S::IRX9-GFP::NosT	EC47811 (R2)	EC16527	EC62055	EC16573	EC41421
EC62700	4×p35S::*At*IRX9-eGFP::NosT	EC47822 (R3)	EC16527	EC62053	EC16570	EC41421
EC62701	4×p35S::*At*IRX9[G28I]-eGFP::NosT	EC47822 (R3)	EC16527	EC62303	EC16570	EC41421
EC62702	4×p35S::*At*IRX9[C24S]-eGFP::NosT	EC47822 (R3)	EC16527	EC62306	EC16570	EC41421
EC62198	4×p35S::*At*IRX10-FLAG::NosT	EC47811 (R2)	EC16527	EC62055	EC16573	EC41421
EC62703	4×p35S::*At*IRX14-HA::NosT	EC47802 (R1)	EC16527	EC62057	EC16572	EC41421

Table 2.6 Level 2 binary vectors assembled in this work

Name	Contents	Backbone	Inserted parts			
pJAC1	pre-assembled Golden Gate backbone for Arabidopsis expression	EC50506	EC62111	EC47811	EC62112	EC41766
EC62555	pIRX14::*At*IRX9-Myc::NosT	pJAC1	EC62312	EC62053	EC16574	EC41421
EC62556	pIRX14::*At*IRX9[G28I]-Myc::NosT	pJAC1	EC62312	EC62303	EC16574	EC41421
EC62557	pIRX14::*At*IRX9[C32S]-Myc::NosT	pJAC1	EC62312	EC62306	EC16574	EC41421
EC62252	4×p35S::*At*IRX9-eGFP::NosT:: 4×p35S::*At*IRX14-HA::NosT 4×p35S::*At*IRX10-FLAG::NosT:: 4×p35S::ManI-mCherry::NosT	EC50506	EC62197	EC62198	EC62199	EC62400 EC41780
EC62254	4×p35S::*At*IRX9-eGFP::NosT:: 4×p35S::ManI-mCherry::NosT	EC50506	EC62197	EC62402	EC41744	
EC62507	4×p35S::ManI-mCherry::NosT	EC50506	EC62197	EC41722		
EC62521	4×p35S::*At*IRX9-eGFP::NosT:: 4×p35S::*At*IRX10-FLAG::NosT:: 4×p35S::*At*IRX14-HA::NosT	EC50506	EC62703	EC62198	EC62700	EC41766
EC62522	4×p35S::*At*IRX9[G28I]-eGFP::NosT:: 4×p35S::*At*IRX10-FLAG::NosT:: 4×p35S::*At*IRX14-HA::NosT	EC50506	EC62703	EC62198	EC62701	EC41766
EC62523	4×p35S::*At*IRX9[C24S]-eGFP::NosT:: 4×p35S::*At*IRX10-FLAG::NosT:: 4×p35S::*At*IRX14-HA::NosT	EC50506	EC62703	EC62198	EC62702	EC41766
EC62534	pIRX3::*Eg*XAPT[A235G]-Myc::NosT	pJAC1	EC62016	EC62341	EC16574	EC41421
EC62506	pIRX3::FVE16326-Myc::NosT	pJAC1	EC62016	EC62326	EC16574	EC41421
EC62526	pXXT2::Bv7_174350-Myc::NosT	pJAC1	EC62311	EC62334	EC16574	EC41421
EC62528	pXXT2::*Vm*GT47-A12-Myc::NosT	pJAC1	EC62311	EC62336	EC16574	EC41421
EC62529	pXXT2::Cc07_g06550-Myc::NosT	pJAC1	EC62311	EC62337	EC16574	EC41421
EC62530	pXXT2::Cc07_g06570-Myc::NosT	pJAC1	EC62311	EC62338	EC16574	EC41421
EC62552	pCESA3::*Vm*GT47-A12-Myc::NosT	pJAC1	EC62031	EC62336	EC16574	EC41421
EC62553	pCESA3::Cc07_g06550-Myc::NosT	pJAC1	EC62031	EC62337	EC16574	EC41421
EC62566	pCESA3::*Sl*XST1-Myc::NosT	pJAC1	EC62031	PCR product	EC16574	EC41421

2.3.2 Site-directed mutagenesis of level 0 MoClo parts

Pairs of divergent primers annealing around the site of mutagenesis were designed; one in each pair contained a small 5' mismatch comprising the mutation to be made (**Table 2.7**). The entire length of the plasmid was then amplified by PCR using Q5® High-Fidelity DNA Polymerase (NEB), with an extension time of 3 min. The PCR product was subjected to electrophoresis in 1 % agarose with 0.001 % SYBR® Safe DNA Gel Stain (Invitrogen) and isolated by gel extraction using the QIAquick Gel Extraction Kit (QIAGEN) followed by isopropanol precipitation (see below). Purified PCR product was simultaneously 5'-phosphorylated and ligated in an overnight room-temperature reaction containing 0.5 µl PCR product diluted 1:20 in PCR-grade water, 1 µl T4 DNA ligase buffer (NEB), 1 µl T4 polynucleotide kinase (NEB), 1 µl T4 DNA ligase (NEB), and 15.5 µl PCR-grade water. Ten microlitres of the ligation reaction was then transformed into *E. coli* 5-alpha (NEB). After purification with the Plasmid Miniprep Kit (QIAGEN), correct mutagenesis was verified by sequencing by the DNA Sequencing Facility (Department of Biochemistry, University of Cambridge).

Table 2.7 Primers used for site-directed mutagenesis.

Mutation	Forward primer (5'→3')	Reverse primer (5'→3')
IRX9[G28I]	ATTTTCTTCACTGGCTTCGCTCCTG	CATTACAAAACATAGAGAGAAATGGATCAC
IRX9[C24S]	CTTTTGTAATGGGTTTCTTCACTGGC	ATAGAGAGAAATGGATCACAGCTTTC
*Eg*XAPT[A235G]	GATGGGACTTCCTCCGCCTCTCG	CGGTCCGGCCGAGGACCAG

2.3.3 DNA extraction from plant material

For Arabidopsis and *Solanum lycopersicum*, a ~1 mm^2 section of leaf was mashed using a pipette tip in 20 µl weak TNE/SDS buffer (20 mM Tris-HCl, *p*H 8.0, 12.5 mM NaCl, 2.5 mM EDTA, 0.05 % SDS) before freezing at −20 °C. After defrosting, 0.5 µl of the supernatant was used as a template for PCR reactions.

For Myrtaceæ-family plants, a CTAB method was used. A handful of leaves was ground to a fine powder in liquid nitrogen using a pre-cooled mortar and pestle. One millilitre of sorbitol wash buffer (100 mM Tris-HCl, *p*H 8.0, 0.35 M sorbitol, 5 mM EDTA, 3 % polyvinylpyrrolidone, 10 mM DTT), preheated to 60 °C, was added to approximately 200 µl of powder from fresh leaves or 50 µl from lyophilised leaves before incubation at 60 °C for 30–40 min with regular inversion. Samples then underwent centrifugation at 3,000 × *g* for 10 min; the supernatant was discarded and the sorbitol wash step was repeated with a shorter

incubation of 2–5 min. After a second round of centrifugation, 800 µl CTAB extraction buffer (100 mM Tris-HCl, *p*H 8.0, 2 % cetyltrimethylammonium bromide, 1.5 M NaCl, 20 mM EDTA, 3 % polyvinylpyrrolidone, 10 mM DTT), pre-heated to 60 °C, was added to the pellet. The sample was mixed by shaking before a final incubation at 60 °C for 30 min with regular inversion. The sample then underwent centrifugation at 10,000 × *g* for 15 min. The supernatant was carefully transferred to a new tube using a wide-bore pipette before adding 1 ml 24:1 chloroform:isoamyl alcohol and mixing by inversion. To separate the phases, this sample underwent another centrifugation at 10,000 × *g* for 10 min. The aqueous phase (~600 µl) was carefully transferred to a new tube and the DNA isolated by isopropanol precipitation.

2.3.4 Purification of DNA by isopropanol precipitation

For one volume of DNA solution, a tenth of a volume of 3 M sodium acetate, *p*H 5.2, was added, followed by one volume of cold isopropanol. The tube was mixed by inversion and incubated at −20 °C for 1–18 h before centrifugation at 20,000 × *g* at 4 °C for 30 min. The pellet was carefully washed with cold 70 % ethanol before a second centrifugation. The supernatant was carefully removed and the pellet left to air dry before resuspending in an appropriate amount of PCR-grade water, incubating at 37 °C for 30 min to aid resuspension where necessary.

2.3.5 Sequencing of unmapped plant genomic DNA using PCR

All PCR reactions were conducted using Q5® High-Fidelity DNA Polymerase (NEB) according to manufacturer's recommendations, except that 40 cycles of amplification were typically used. For PCR amplification of XAPT1 homologues (but not XLPT1 homologues), the GC-rich enhancer reagent was added. Approximately 10–100 ng of genomic DNA was used in each reaction. PCR products were separated by electrophoresis in 1 % agarose with 0.001 % SYBR® Safe DNA Gel Stain (Invitrogen) and isolated by gel extraction using the QIAquick Gel Extraction Kit (QIAGEN) followed by isopropanol precipitation. In the case of poor yield, extracted PCR products became templates for a second round of amplification. PCR products were sequenced by the DNA Sequencing Facility (Department of Biochemistry, University of Cambridge) using the same primers as for amplification.

2.3.6 Bacterial transformation

E. coli was transformed by conventional heat-shock transformation. Briefly, plasmid DNA was added to 40 µl competent cells and incubated for 30 mins before heat shock at 42 °C for 30 s. Cells were cooled immediately on ice for 2 min before the addition of 500 µl LB medium and

incubation at 37 °C for 1 h. Cells were pelleted, resuspended in a small volume of LB, and spread on an LB + agar plate supplemented with the appropriate antibiotic.

For *Agrobacterium tumefaciens* AGL1 and GV3101 strains, 250–1000 ng plasmid DNA was added to each 100 µl competent cell aliquot (gifts of Henry Temple and Yoshihisa Yoshimi). The cells were then frozen instantaneously in liquid nitrogen before thawing at 37 °C: 5 min for AGL1 and 3 min for GV3101. One millilitre of LB medium was then added and the cells incubated at 30 °C for 3 h with gentle shaking. The cells were then pelleted at 10,000 × *g* for 2 min before resuspension in a small volume of LB and spreading on an LB + agar plate supplemented with antibiotics for the selection of both the strain-intrinsic helper plasmid (100 µg ml^{-1} carbenicillin for AGL1 or 50 µg ml^{-1} gentamicin for GV3101) and the newly transformed plasmid.

2.3.7 Plant transformation

Arabidopsis plants were transformed using the floral dip protocol as described by Clough & Bent (1998). *A. tumefaciens* GV3101 was used to transform 6–7-week old plants; 50 ml of stationary or near-stationary phase culture was pelleted and resuspended in an equal volume of dipping medium. A cassette for the seed-specific expression of GFP-tagged oleosin (Shimada *et al*, 2010) was placed in all Arabidopsis transformation constructs; hence, transgenic seeds was selected by virtue of their GFP fluorescence, which was observed using a stereo microscope fitted with a GFP fluorescence illuminator. In later generations, zygosity could be screened by observing ratios of fluorescent:non-fluorescent seeds in siliques.

Transient expression in tobacco leaves was performed essentially as described by Sparkes *et al* (2006). Briefly: 25 ml *A. tumefaciens* AGL1 cultures were grown to an OD_{600} of 0.6–1 on the day of infiltration. Cells were pelleted and resuspended in 25 ml infiltration medium. The cell volumes of each were equalised according to the OD_{600} by discarding a proportional volume of the suspension. Cells were pelleted again and resuspended in 12.5 ml infiltration medium. The IRX culture suspensions were then each mixed with an equal volume of ManI-mCherry culture suspension to permit co-infiltration. The youngest non-wrinkled leaves from four-week old *Nicotiana benthamiana* plants were infiltrated from the abaxial side using a 1 ml blunt syringe body.

2.3.8 TOXGREEN experiments

Pairs of complementary oligonucleotides constituting the TMH-encoding plasmid inserts were synthesised and PAGE-purified by Sigma-Aldrich (**Table 2.8**). These oligonucleotides were

designed either to contain *Dpn*II and *Nhe*I cut sites at their termini or such that annealing would produce *Nhe*I- and *Bam*HI-compatible overhangs at the 5' and 3' termini respectively. Pairs were then annealed at 25 µM of each oligonucleotide in 1× T4 DNA ligase buffer (NEB) by heating at ~99 °C for 5 min before cooling on ice. For those that required digestion, oligonucleotides were first cut with *Nhe*I (NEB) according to manufacturer's recommendations, purified with the QIAquick PCR purification kit (QIAGEN), then cut with *Dpn*II (NEB) according to the manufacturer's recommendations before a second round of purification (this approach required significant amounts of starting product to compensate for losses during purification). Oligonucleotides with overhangs produced by either method were then 5'-phosphorylated in a reaction containing 4 µl ~0.15 µM oligonucleotide, 1 µl 10× T4 DNA ligase buffer (NEB), 0.5 µl T4 polynucleotide kinase (NEB), and 4.5 µl PCR-grade water, which was heated at 37 °C for 30 min and then 65 °C for 20 min. The backbone acceptor, pccGFPKAN, was a gift from Alessandro Senes (Addgene plasmid # 73649). Purified pccGFPKAN was cut and dephosphorylated in a reaction containing ~2.5 µg pccGFPKAN, 3 µl *Bam*HI-HF® (NEB), 3 µl *NheI* (NEB), 5 µl CIP alkaline phosphatase (NEB), 30 µl 10× CutSmart® buffer, and PCR-grade water to a total volume of 300 µl. This reaction was incubated at 37 °C for 15 mins before isolation of the DNA using the QIAquick PCR purification kit (QIAGEN). Each insert was then ligated into the backbone acceptor in a reaction containing ~30 ng backbone, 0.5 ng insert, 2 µl 10× T4 DNA ligase buffer (NEB), 1 µl T4 DNA ligase (NEB), and PCR-grade water to a total volume of 20 µl, which was incubated at 16 °C overnight before termination at 65 °C for 10 min. A 10 µl aliquot of the ligation reaction was then transformed into competent *E. coli* 5-alpha (NEB). Plasmids were purified using the Plasmid Miniprep Kit (QIAGEN); correct assembly was verified by sequencing by the DNA Sequencing Facility (Department of Biochemistry, University of Cambridge).

Table 2.8 Complementary oligonucleotide pairs used to create transmembrane inserts in pccGFP plasmids.

Name	Sense strand sequence (5'→3')	Antisense strand sequence (5'→3')
GpA	GGGGCTAGCCTCATTATTTTTGGGGTGATGGCTGGCGTTATTGGAACGATCGGGG	CCCCGATCGTTCCAATAACGCCAGCCATCACCCCAAAAATAATGAGGCTAGCCCC
GpA*	GGGGCTAGCCTCATTATTTTTGGGGTGATGGCTATTGTTATTGGAACGATCGGGG	CCCCGATCGTTCCAATAACAATAGCCATCACCCCAAAAATAATGAGGCTAGCCCC
IRX9-1 ($IRX9_{17–36}$)	GGGGCTAGCGCGGTGATTCATTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGCGCCCGGCGATCGGGG	CCCCGATCGCCGGCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAATGAATCACCGCGCTAGCCCC
IRX9-2 ($IRX9_{14–36}$)	GGGGCTAGCTGGAAGAAAGCGGTGATTCATTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGCGCCCGGCGATCGGGG	CCCCGATCGCCGGCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAATGAATCACCGCTTTCTTCCAGCTAGCCCC
IRX9-3 ($IRX9_{14–34}$)	GGGGCTAGCTGGAAGAAAGCGGTGATTCATTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGCGATCGGGG	CCCCGATCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAATGAATCACCGCTTTCTTCCAGCTAGCCCC
IRX9-4 ($IRX9_{17–34}$)	CTAGCGCGGTGATTCATTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGC	GATCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAATGAATCACCGCG
IRX9-5 ($IRX9_{21–34}$)	CTAGCTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGC	GATCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAG
IRX9-5m ($IRX9_{21–34}$*)	CTAGCTTTAGCCTATGCTTTGTGATGATTTTTTTTACCGGCTTTGC	GATCGCAAAGCCGGTAAAAAAAATCATCACAAAGCACAGGCTAAAG
IRX9-6 ($IRX9_{21–37}$)	CTAGCTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGCGCCGGCGGG	GATCCCCGCCGGCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAG
IRX9-7 ($IRX9_{20–37}$)	CTAGCCATTTTAGCCTGTGCTTTGTGATGGGCTTTTTTACCGGCTTTGC	GATCGCAAAGCCGGTAAAAAAGCCCATCACAAAGCACAGGCTAAAATGG
IRX14-1 ($IRX14_{35–59}$)	GGGGCTAGCATTGCGGTGTTTTGGCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGGCTTTCGCTTTTCGATCGGGG	CCCCGATCGAAAAGCGAAAGCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGCCAAAACACCGCAATGCTAGCCCC
IRX14-2 ($IRX9_{40–62}$)	GGGGCTAGCTGGCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGGCTTTCGCTTTAGCCGCCTGGTGATCGGGG	CCCCGATCACCAGGCGGCTAAAGCGAAAGCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGCCAGCTAGCCCC
IRX14-3 ($IRX9_{40–59}$)	GGGGCTAGCTGGCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGGCTTTCGCTTTTCGATCGGGG	CCCCGATCGAAAAGCGAAAGCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGCCAGCTAGCCCC
IRX14-4 ($IRX9_{40–66}$)	GGGGCTAGCTGGCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGGCTTTCGCTTTAGCCGCCTGGTGTTTTTTTTTCTGATCGGGG	CCCCGATCAGAAAAAAAAACACCAGGCGGCTAAAGCGAAAGCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGCCAGCTAGCCCC
IRX14-5 ($IRX9_{42–59}$)	CTAGCATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGGCTTTCGCTTTTC	GATCGAAAAGCGAAAGCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATG
IRX14-6 ($IRX9_{40–55}$)	CTAGCTGGCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGG	GATCCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGCCAG
IRX14-7 ($IRX9_{36–55}$)	CTAGCATTGCGGTGTTTTGGCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGG	GATCCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGCCAAAACACCGCAATG
IRX14-8 ($IRX14_{41–55}$)	CTAGCCTGATTCTGCATTGCCTGTGCTGCCTGATTAGCCTGGTGCTGGG	GATCCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATCAGG
IRX14-9 ($IRX9_{42–55}$)	CTAGCATTCTGCATTGCCTGTGCTGCCTCATTAGCCTGGTGCTGGG	GATCCCCAGCACCAGGCTAATCAGGCAGCACAGGCAATGCAGAATG

E. coli MM39 was a gift of Bryan Berger (Addgene strain # 42894). TOXGREEN pccGFP plasmids were transformed into competent cells prepared as described by Im *et al* (2011). For maltose uptake complementation tests, a single colony was picked and streaked across solid M9 minimal medium supplemented with 0.4 % maltose, prepared as described by Armstrong & Senes (2016), and incubated at 37 °C for 72 h. For GFP fluorescence measurements, three separate colonies were picked for each construct and used to inoculate 3 ml of LB medium with 100 µg ml^{-1} carbenicillin. These cultures were incubated at 37 °C overnight to reach stationary phase. From each culture, two 300 µl aliquots were loaded into either a transparent or a black 96-well plate, which, using a BMG Labtech FLUOstar OPTIMA Fluorescence Microplate Reader, permitted measurement of OD_{590} and 485/520 nm fluorescence, respectively. Fluorescence readings were adjusted according to the OD_{590} reading in a linear fashion.

For Western blotting, 100-µl culture aliquots were pelleted at maximum speed in a benchtop microcentrifuge. The supernatant was removed before adding 30 µl SDS-PAGE buffer containing 0.625 M Tris-HCl, *p*H 6.8, 6.25 % glycerol, 0.01 % bromophenol blue, 2 % SDS, and 100 mM DTT. This solution was boiled at 100 °C for 10 min before pelleting. For these samples, 5 µl aliquots were loaded into a 10-well 4–15% Mini-PROTEAN® TGX™ precast protein gel (Bio-Rad), and a potential difference of 100 V was applied for 30 min, then 150 V for 45 min. Protein was transferred to a nitrocellulose membrane using the iBlot Dry Blotting System (Invitrogen). Western blotting proceeded by the protocol in **Table 2.9**; the primary antibody was 1:10000 rabbit polyclonal anti-MBP (Abcam; ab9084) and the secondary was 1:10000 HRP-linked goat anti-rabbit (Bio-Rad). The blot was treated with Amersham ECL Prime Western Blotting Detection Reagent (GE healthcare) and imaged using Amersham Hyperfilm ECL (GE healthcare).

Table 2.9 Western blotting steps. TBS is 100 mM Tris-HCl, pH 7.4 with 500 mM NaCl; TBST is TBS with 0.1 % Tween-20.

Step	Solution	Duration
Membrane blocking	5 % milk in 1× TBS	2 h or overnight
Primary antibody	Antibody diluted in blocking solution	2 h
Wash	1× TBST	3 × 10 min
Secondary antibody	Antibody diluted in blocking solution	1 h
Wash	1× TBS	5 × 5 min

2.4 Protein purification

Unpurified recombinant human EXTL3 protein (Δ1-51; N-terminally His-tagged), expressed in EBNA 293 cells, collected in EX-CELL® 325 PF CHO Serum-Free Medium (Merck), and dialysed against a buffer containing 50 mM sodium phosphate, *p*H 7.0, and 0.3 M NaCl, was a gift of Katrin Man—see Awad *et al* (2018). Protein was purified with assistance from Steven Hardwick in a weak buffer containing 50 mM Tris-HCl, *p*H 6.8, 100 mM NaCl, and 50 mM KCl by nickel affinity chromatography using a 1 ml HisTrap HP (GE Healthcare) column with a linear imidazole gradient (0→250 mM) followed by gel filtration chromatography using a Superdex™ 200 Increase 10/300 GL column (GE Healthcare). Protein was concentrated to ~0.1–6.0 mg ml^{-1} using a VivaSpin 500 30 kDa concentrator (Vivascience).

2.5 Glycan manipulation and analysis

2.5.1 Alcohol-insoluble residue (AIR) preparation

Plant material was homogenised in absolute ethanol using a Retsch MM400 ball mill at 25 rpm with a 10/5 min on/off cycle (for the more resistant samples this was changed to 30 rpm 5/5 min in later cycles). For Myrtaceæ plants, tissues were separated by dissection prior to milling: the outer stem (comprising the epidermis, cortex, and probably the phloem) was separated from the core (comprising the pith and xylem) by peeling or scraping. The milled material was washed in absolute ethanol after centrifugation at 3,000 × *g* for 10 min. After further centrifugation the pellet was resuspended in 2:3 methanol:chloroform and incubated in a shaker overnight. This material was treated a second time in fresh methanol:chloroform for one hour before successive washes in absolute, 65 %, and 80 % ethanol, followed by another wash in absolute ethanol. The supernatant was removed after the final wash and the remaining material (alcohol-insoluble material; AIR) was dried for three days in an oven at 45 °C. *Eucalyptus dalrympleana* leaf, epidermal, cortical, and core stem AIR was a gift of Li Yu (University of Cambridge); *Solanum lycopersicum* fruit AIR (incl. depectinated material) was a gift of Yoshihisa Yoshimi (University of Cambridge). *Plinia cauliflora*, *Psidium guajava*, and *Myrcianthes pungens* stem material was a gift of Pedro Araújo (University of Campinas). Plant material for *Myrtus communis*, *Metrosideros excelsa*, and *Coffea arabica* was provided by the Cambridge University Botanic Garden. *Vaccinium corymbosum* (blueberry) fruits were purchased from Sainsbury's supermarket (42–45 Sidney Street, Cambridge, UK). *Artemisia dracunculus* (tarragon) leaves were purchased from Tesco Superstore (Cheddars Ln, Cambridge, UK). *Psidium guajava* (guava) fruit was purchased from Spice Gate (14 Mill Rd, Cambridge, UK).

2.5.2 Monosaccharide analysis

Monosaccharide analysis was performed as described by Bromley *et al* (2013). Briefly, 50 µg AIR was resuspended in 400 µl 2 M trifluoroacetic acid and heated at 120 °C for 1 h. Samples were then dried in a centrifugal evaporator and analysed alongside 500 pmol monosaccharide standards using a Dionex ICS300 HPLC system. Data were processed with Chromeleon software (Dionex).

2.5.3 Extraction and enzymatic digestion of xylan

To extract xylan, 0.25–0.5 mg AIR was resuspended in 20 µl 4 M NaOH and incubated at room temperature for 1 h before neutralisation with 80 µl 1 M HCl. Nine hundred microlitres of 50 mM ammonium acetate, *p*H 5.5–6.0 was then added.

For GH11 xylanase digestions, 3 µl 10 kU ml^{-1} GH11 from *Neocallimastix patriciarum* (E-XYLNP, Megazyme) was added to 500 µl of the above and the reaction incubated at 30 °C overnight with regular periods of vigorous shaking. For GH30 xylanase digestions, 1 µl 2.7 mg ml^{-1} XynA from *Dickeya (Erwinia) chrysanthemi* P860219 (GenBank: AAL16415.1; this particular enzyme being able to accommodate α1,2-arabinopyranosylated glucuronic acid), a gift of Oliver Terrett (formerly University of Cambridge, now moved to ETH Zürich), was added instead, and the reaction was incubated at 25 °C for 40 min with regular vigorous shaking. For double digestions with GH115 α-glucuronidase, 10 µl of 0.1 mg ml^{-1} *Talaromyces leycettanus* GH115 α-glucuronidase (Novozymes) was added either simultaneously with XynA (for *Np*GH11+*Tl*GH115 digestion), or added after terminating the first reaction at 120 °C for 10 min and incubating at 30 °C overnight (for $Ec_{P860219}$GH30+*Tl*GH115 digestion).

For β-D-galactosidase sensitivity assays, alkali-extracted xylan from 0.25 mg AIR was digested in 1 ml of 50 mM ammonium acetate, *p*H 6.0, with 3 µl *Np*GH11 and 5 µl *Tl*GH115 for 18 h at 37 °C. After drying digestions in a centrifugal evaporator, they were resuspended in 500 µl water and passed through a 3 kDa NanoSep centrifugal device (Pall). Samples were dried once more and resuspended in 65 % ethanol before incubation at −20 °C for at least 2 h. After spinning for 10 min at maximum speed in a benchtop microcentrifuge, the supernatant was dried again. The samples were then resuspended in 200 µl 50 mM ammonium acetate, *p*H 5.0. Each was split in two, before the addition of 0.5 µl of 4 kU ml^{-1} GH35 β-galactosidase from *Aspergillus niger* (E-BGLAN, Megazyme) to one 100 µl aliquot in each pair. Reactions were then incubated at 37 °C for 18 h before drying.

2.5.4 Extraction and enzymatic digestion of xyloglucan

To extract xyloglucan, 30 mg AIR was resuspended in 2.7 ml 4 M NaOH and incubated at room temperature for 1 h on a rocker. After centrifugation for 10 min at 3,000 × *g*, 2.5 ml of supernatant was loaded onto a PD-10 desalting column (GE healthcare) pre-equilibrated with 50 mM ammonium acetate, *p*H 6.0. Elution was performed with 3.5 ml of the same buffer. For screens of *xlt2 mur3-1* lines, the PD-10 clean-up was omitted, and xyloglucan was extracted in the same way as xylan (see above).

For xyloglucanase digestions, 930 ul of the same ammonium acetate buffer was added to 70 ul (or equivalent) of the PD-10 eluate. Subsequently, 3 µl 0.6 mg ml^{-1} XG5 xyloglucanase (Novozymes) (Simmons *et al*, 2017) was added before incubation at 37 °C for 18 h. Where necessary, xyloglucanase was removed by resuspending the dry products in 65 % ethanol, incubating at −20 °C for at least 2 h, centrifuging at max speed in a benchtop centrifuge for 10 min, and retaining the supernatant. After drying, these samples were resuspended in 1 ml ammonium acetate buffer.

For *exo*-glycosidase digestions, 100 µl of the 1 ml xyloglucanase products (equivalent to the products from 0.06 mg of the original AIR) was treated with one or more of the following enzymes: 0.5 µl 9.5 µM GH95 α1,2-fucosidase AfcA from *Bifidobacterium bifidum* (Katayama *et al*, 2004) (GenBank: AAQ72464.1; Novozymes), 1 µl lactase/β-galactosidase from *Meripilus* sp. (Novozymes), 1 µl 31 U ml^{-1} GH31 α-xylosidase from *Escherichia coli* (E-AXSEC, Megazyme), 1 µl 40 U ml^{-1} GH3 β-glucosidase from *Aspergillus niger* (E-BGLUC, Megazyme), 1–3 µl 20 µM GH51 α-arabinofuranosidase Abf51 from *Cellvibrio japonicus* (Beylot *et al*, 2001) (GenBank: AAK84947.1; a gift of Harry Gilbert), and/or 2 µl 8 mg ml^{-1} GH3 β1,2-xylosidase from *Chætomium globosum* (Tryfona *et al*, 2019) (NS39127, Novozymes). Reactions were carried out at 37 °C for 18 h, except in the case of *Cg*GH3, for which reactions were carried out over six days. Ethanol precipitations to remove *exo*-glycosidases were as described above for removal of xyloglucanase. An aliquot of 5 µl of the final reaction products was reserved for MALDI-TOF MS; otherwise, the whole sample was dried for subsequent PACE analysis.

2.5.5 Preparation of K5 heparosan oligosaccharides

DP10 (unsaturated) K5 oligosaccharide, prepared by lyase treatment of *E. coli* K5 capsular polysaccharide, was a gift of Katrin Mani (Lund University). DP9 K5 oligosaccharide was prepared by treatment with $BT4658^{GH88}$, a GH88-family Δ-4,5-unsaturated β-glucuronyl

hydrolase from *Bacteroides thetaiotamicron* VPI-5482 (Cartmell *et al*, 2017) (GenBank: AAO79763.1), which was expressed in and purified from *E. coli* BL21(DE3) by Clelton Santos (the BT4658^{GH88} expression plasmid was a gift of Didier Ndeh, Newcastle University). Ten microlitres of ~15 mM DP10 (unsat.) K5 oligo was combined with 12.5 µl 0.1 M ammonium acetate, *p*H 5.5, and 2.5 µl 3 mg ml^{-1} BT4658^{GH88}, and incubated at 30 °C overnight before oligosaccharide isolation using a 3 kDa Nanosep® centrifugal filter (Pall). The effluent was dried *in vacuo* and resuspended in 100 µl dH_2O to create the DP9 stock. DP8 K5 oligosaccharide was prepared from the DP9 by combining 10 µl DP9 stock with 9 µl 50 mM ammonium acetate, *p*H 5.5, and 1 µl 4 mg ml^{-1} fungal *Pa*GH89 α-*N*-acetylglucosaminidase (a gift of Kristian Krogh, Novozymes) and incubated at 30 °C overnight before a second 3 kDa Nanosep® filtration. The sample was dried and resuspended in 10 µl dH_2O. To prepare DP7 K5 oligosaccharide, 2.5 µl DP8 product was combined with 6.5 µl dH_2O, 10 µl 50 mM ammonium acetate, *p*H 5.5, and 1 µl of either 0.8 mg ml^{-1} *Thar*GH79a (THAR02_03122; GenBank: KKP04785.1) or 0.7 mg ml^{-1} *Thar*GH79b (GH79 β-glucuronidases from *Trichoderma harzianum*) and incubated at 30 °C overnight. *Thar*GH79a and *Thar*GH79b were gifts of Kristian Krogh (Novozymes).

2.5.6 Heparosan extension assays

Reactions contained 25 mM ammonium acetate, *p*H 6.5, 12.5 % v/v (~0.1–0.2 mM) stock DP8/DP9 (or DP10 unsat.) K5 oligosaccharide, 0.1 mg ml^{-1} purified human EXTL3, 1.5 mM UDP-GlcA/UDP-GlcNAc, and 3 mM $MnCl_2$/$MgCl_2$ or 10 mM EDTA, and were incubated at 37 °C for 18 h before termination at 100 °C for 10 min (except where indicated). To remove terminal glucuronic acid residues from the products, samples were dried *in vacuo* and resuspended in the same volume of 50 mM ammonium acetate, *p*H 5.5, with either 0.1 mg ml^{-1} bovine liver β-glucuronidase B-1 (Merck) or 0.04 mg ml^{-1} *Thar*GH79a and incubated at 37 °C overnight. For PACE and mass spectrometry analysis, samples were subsequently chemically derivatised (see *Section 2.5.8*). For the UDP detection assay, 'Ultra Pure' UDP-sugars (Promega) were used at the indicated concentrations. Reactions were transferred to nonadjacent wells of a white, opaque 384-well plate, and UDP concentrations were determined using the UDP-Glo™ Glycosyltransferase Assay kit (Promega). Luminescence was detected using a FlexStation 3 Multi-Mode Microplate Reader (Molecular Devices).

2.5.7 Polysaccharide analysis by carbohydrate electrophoresis (PACE)

PACE is an electrophoretic technique for the analytic separation of oligosaccharides (Goubet *et al*, 2002). Very similarly to protein electrophoresis, analytes are loaded into the wells of a

polyacrylamide gel and drawn into and through the gel by an applied potential difference. Oligosaccharides are chemically derivatised at the reducing terminus with a fluorescent moiety to permit detection under UV light; this moiety often also imparts charge. Since some glycans contain anionic monosaccharides, the migration of analytes is affected not only by their size, but also by their native charge. The migration of oligosaccharides through the gel is further influenced by their interaction with borate ions in the alkaline running buffer.

For xylan and xyloglucan digestion products, reactions were dried *in vacuo* and resuspended in either 5 µl or 20 µl (depending on sample amount) of labelling reagent comprising 50 % DMSO, 7.5 % acetic acid, 50 mM 2-picoline-borane (2-PB), and 50 mM 8-aminonaphthalene-1,3,6-trisulphonic acid (ANTS), before incubating at 37 °C overnight in the dark. Samples were dried again and resuspended in 10–50 µl 6 M urea. For EXTL3 heparosan extension reactions, a 5 µl aliquot (or a 1.25 µl aliquot of DP8/DP9 preparations) was dried and resuspended in 10 µl of labelling reagent with higher reducing potential, comprising 25 % DMSO, 3.75 % acetic acid, at least 0.23 M 2-PB, and 50 mM ANTS. Samples were incubated at 37 °C overnight, dried and resuspended in 5 µl 6 M urea.

PACE was carried out as previously described (Goubet *et al*, 2002). A 2.5 µl aliquot of each sample (as well as of labelled standards) was loaded into the 20 % polyacrylamide gel. For heparosan extension assays, the gel underwent 200 V for 30 min, then 1000 V for 2 h 10 min to maximise separation; similarly, gels for xyloglucan experiments underwent 200 V for 30 min, then 1000 V for 2 h 30 min.

2.5.8 Mass spectrometry

For heparosan extension assays, completed reactions (95 µl volume) were filtered using a 10 kDa Nanosep® centrifugal filter (Pall). The filtrate was dried *in vacuo* before desalting by cation exchange using a column packed with Dowex® beads (50×8, H^+ form, 50–100 mesh; Merck) as described by Tryfona & Stephens (2010). The eluate was dried again before reducing-end derivatisation with 2-aminobenzamide (2-AB) with subsequent purification using a GlycoClean™ S Cartridge (ProZyme) as described by Tryfona & Stephens (2010). The eluate was dried and resuspended in dH_2O. Sample spotting and mass spectrometry (in negative mode) were then carried out by Theodora Tryfona.

For analysis of xyloglucan-derived oligosaccharides, underivatized products (1 µl undiluted and tenfold-diluted) were spotted directly onto the mass spectrometry plate and overlaid with 1 µl 130 mM 2,5-dihydroxybenzoic acid and 3 mM ammonium sulphate in 50 % methanol.

The mass spectrometer was operated (in positive mode) by Li Yu. Spectra were analysed using Bruker Daltonics flexAnalysis software.

2.6 Confocal and cryo-electron (cryo-EM) microscopy

2.6.1 Confocal laser scanning microscopy

Three days following infiltration, small ~0.5 cm^2 sections were cut from tobacco leaves transiently expressing fluorescently tagged proteins of interest. These sections (at least three per construct) were placed in water in between the slide and coverslip with the abaxial side facing upwards. Confocal images for the initial IRX9 *vs* IRX9, IRX10, and IRX14 experiment were acquired using a Leica SP8 laser scanning confocal microscope with a 63× water objective, which was operated by Henry Temple. GFP and mCherry were excited using 488 and 552 nm-wavelength lasers, respectively. Confocal images for IRX9 mutants were acquired using a Zeiss LSM 880 laser scanning confocal microscope with a 63× water objective. GFP and mCherry were excited using 488 and 561 nm-wavelength lasers, respectively. The previously characterised Golgi marker constituted the first 59 amino acids of a mannosidase I variant from *Glycine max* (Nelson *et al*, 2007; Lao *et al*, 2014).

2.6.2 Preparation of cryo-EM grids

For *apo*-EXTL3, QUANTIFOIL® R 1.2/1.3 holey carbon grids were glow-discharged once using the PELCO easiGlow™ system (Ted Pella). Three microlitres of ~0.1 mg ml^{-1} freshly purified EXTL3 were applied to the grid, which was blotted and frozen in liquid ethane using a Vitrobot Mark IV (FEI) system with blot force 0, blot time 3 s, and at 4 °C and 95 % humidity by Steven Hardwick.

For EXTL3 with UDP and Mn^{2+}, QUANTIFOIL® R 2/2 holey carbon grids were glow-discharged once on each side. Purified EXTL3, stored at −80 °C with 20 % glycerol, was dialysed against 50 mM Tris-HCl, *p*H 6.8, 100 mM NaCl, and 50 mM KCl over the preceding night. Two and a half hours prior to grid freezing, EXTL3 concentration was estimated using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher), and a solution comprising 50 mM Tris-HCl, *p*H 6.8, 100 mM NaCl, 50 mM KCl, 2.5 mM $MnCl_2$, ~0.4 mg ml^{-1} EXTL3, and 10 mM UDP was combined and placed on ice. Three microlitres of this solution were applied to the grid, which was blotted and frozen in liquid ethane using a Vitrobot Mark IV (FEI) system with blot force +2, blot time 3 s, and at 4 °C and 95 % humidity.

Grids were stored under liquid nitrogen.

2.6.3 Cryo-EM data collection

Grids were screened using a Talos Arctica™ (Thermo Fisher) instrument by Steven Hardwick and Tom Dendooven. Data was acquired using a Titan Krios™ (Thermo Fisher) instrument by Dima Chirgadze. Data collection parameters are listed in **Table 2.10**.

Table 2.10 Cryo-EM data collection parameters

	apo-EXTL3	**UDP-bound EXTL3**
Detector	Falcon III	Gatan K2
Magnification (nominal)	120k	130k
Energy filter slit size (eV)	none	20
Voltage (kV)	300	300
Total electron dose (e^{-}/Å^2)	71.4	48.5
Target defocus range (μm)	−3.0 – −0.9	−2.7 – −1.2
Calibrated pixel size (Å)	0.667	1.065
Movies collected	1,440	2,573

2.6.4 Cryo-EM data processing

Apo-EXTL3 data was processed in RELION-2.1 (Scheres, 2012; Kimanius *et al*, 2016) with assistance from Tom Dendooven. Beam-induced motion correction was achieved with MotionCor2 (Zheng *et al*, 2017) and the contrast transfer function was estimated using Gctf1.06 (Zhang, 2016). A total of 656,292 particles were auto-picked and extracted with 3× binning. These particles then underwent several rounds of 2D classification before generation of a 3D initial model with C1 symmetry. Subsequently, 3D classification was performed. The resultant 171,285 particles were re-extracted without binning and with a box size of 420 px. A new initial 3D model and mask with C2 symmetry were generated using cryoSPARC™ before 3D refinement. After movie refinement and particle polishing, the final 3D refinement and post-processing steps were performed at a pixel size of 0.67 Å, producing a map at 2.9 Å resolution (criterion: Fourier shell correlation (FSC) = 0.143). Pixel size calibration was achieved by fitting of the GT64 domain portion of the map to a single chain from the

crystallographic structure of mouse EXTL2 (PDB: 1OMX). The map was deposited in the Electron Microscopy Data Bank under accession 11923.

Ab initio modelling was achieved with Buccaneer (Hoh *et al*, 2020) via the CCP-EM interface (Burnley *et al*, 2017). The model was refined with *Coot* (version 0.8.9.2-pre) (Emsley *et al*, 2010), ISOLDE (Croll, 2018), and Phenix real-space refine (Afonine *et al*, 2013). Data processing and refinement parameters are listed in **Table 2.11**. The co-ordinates were deposited in the Protein Data Bank under accession 7AU2.

EXTL3 with UDP and Mn^{2+} was processed in RELION-3 (Zivanov *et al*, 2018). MotionCor2 was used to apply beam-induced motion correction before CTF estimation using Gctf v1.06. Thirty-eight movies with strong ice-derived Thon rings were manually removed; subsequently, 1,133,069 particles were auto-picked and extracted with 3× binning. These particles then underwent several rounds of 2D classification and one round of 3D classification before repeating auto-picking using the best model as a 3D reference. A total of 1,696,344 particles were auto-picked and extracted before further 3D classification, using the same reference. The resultant 238,379 particles were re-extracted without binning and with a box size of 360 px. A new initial 3D model and mask with C2 symmetry were generated using cryoSPARC™ before 3D refinement. After two rounds each of particle polishing and CTF refinement, the final 3D refinement and post-processing steps were performed at a pixel size of 1.06 Å, producing a map at 2.9 Å resolution (criterion: FSC = 0.143). The map was deposited in the Electron Microscopy Data Bank under accession 11926.

To create the UDP-bound EXTL3 co-ordinate model, the apo-EXTL3 model was docked into the map using UCSF Chimera (Pettersen *et al*, 2004) before further refinement using *Coot*, ISOLDE, and Phenix real-space refine. The final co-ordinates were deposited in the Protein Data Bank under accession 7AUA.

Table 2.11 EXTL3 atomic co-ordinates: refinement parameters

	apo-EXTL3	UDP-bound EXTL3
Map resolution at FSC = 0.143 (Å)	2.4	2.9
Model composition		
Non-hydrogen atoms	11,236	11,300
Protein residues	1,372	1,372
Non-protein atoms	156	208
B factors (Å^2)		
Protein	39.55	56.32
Glycans/ligand	46.21	68.81
R.m.s. deviations		
Bond lengths (Å)	0.006	0.008
Bond angles (°)	1.043	0.982
Validation		
MolProbity score	1.06	1.19
Clashscore	2.29	1.84
Poor rotamers (%)	0.83	0
Ramachandran plot		
Favored (%)	97.79	96.31
Allowed (%)	2.21	3.69
Disallowed (%)	0	0

2.7 Thesis typesetting and preparation of figures

Phylogenies were rendered in FigTree. Bar charts were plotted using Microsoft Excel 365; mass spectra were plotted using RStudio. The contrast of colour photographs was improved using ImageJ (Schneider *et al*, 2012). Protein structures and cryo-EM density maps were rendered in Pymol v2.4 (Schrödinger LLC, 2020), UCSF Chimera, and ChimeraX (Pettersen *et al*, 2021). All figures were prepared for inclusion in the manuscript using Inkscape. The manuscript was prepared using Microsoft Word 365.

Chapter 3 : Transmembrane dimerisation of IRX9 and IRX14

3.1 Introduction

As discussed in *Section 1.5.2*, IRX9 and IRX14 (and their paralogues IRX9L and IRX14L) are GT43-family proteins involved in xylan synthesis in Arabidopsis (Anders & Dupree, 2011). Both are proposed to interact with IRX10/10L to form the xylan synthase complex (XSC) (Wierzbicki *et al*, 2019). Unlike IRX14/14L, IRX9/9L is not thought to possess glycosyltransferase activity, and has therefore been proposed to perform a structural role in the assembly (Ren *et al*, 2014). However, the stoichiometry and interactions between the XSC components have not been fully resolved (Wierzbicki *et al*, 2019).

So far, biochemical evidence for the interaction of GT43- and GT47-family proteins to form an XSC has come from two experimental systems—involving wheat and Asparagus IRX orthologues, respectively (Zeng *et al*, 2010; Jiang *et al*, 2016; Song *et al*, 2015; Zeng *et al*, 2016). In the former, an antibody to *Ta*GT43-4 (an IRX14 orthologue) was used to co-immunoprecipitate a protein complex exhibiting xylosyl-, glucuronosyl-, and arabinosyltransferase activities from etiolated wheat seedling microsomes (Zeng *et al*, 2010). Subsequent proteomic analysis suggested that this complex contained *Ta*GT43-4, *Ta*GT47-13 (an IRX10 orthologue), *Ta*GT75-3 and *Ta*GT75-4 (homologues of a characterised UDP-Ara*p* mutase), *Ta*GLP (a 'germin-like protein'), and *Ta*VER2 (a 'vernalisation-like protein'); however, an IRX9 orthologue was not identified in the analysis (Jiang *et al*, 2016). Native-PAGE experiments were able to detect the formation of a higher molecular weight complex when *Ta*GT43-4 and *Ta*GT47-13, or *Ta*GT43-4, *Ta*GT47-13, *Ta*GT75-3, and *Ta*GT75-4 were co-expressed heterologously in the yeast *Pichia pastoris*, but not when *Ta*GT43-4 and *Ta*GT47-13 were expressed separately (Jiang *et al*, 2016). Bimolecular fluorescence complementation (BiFC) experiments on fluorescently tagged proteins expressed transiently in *Nicotiana benthamiana* leaves provided evidence for various interactions between these proteins—as summarised in **Figure 3.1a;** furthermore, the results of these experiments suggested that *Ta*GT43-4, but not *Ta*GT47-13, forms homodimers (Jiang *et al*, 2016). Additional assays in *N. benthamiana* leaves revealed that *Ta*GT43-4 is localised to the ER in the absence of *Ta*GT47-13, whereas *Ta*GT43-4 co-expressed with *Ta*GT47-13 is localised to the Golgi (Jiang *et al*, 2016).

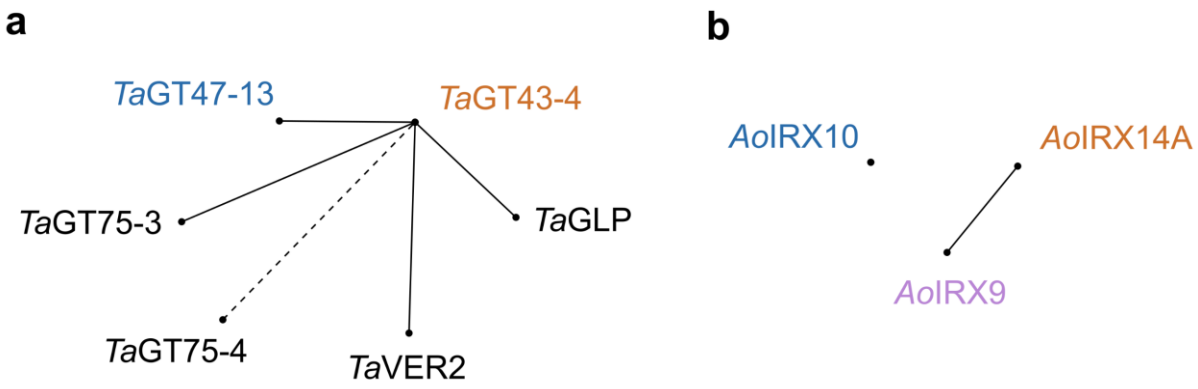


Figure 3.1 Reported interactions between putative XSC members, as determined by bimolecular fluorescence complementation. a Interactions between putative wheat XSC members as reported by Jiang *et al* (2016). The dotted line represents the fact that the interaction was not directly assayed, but rather assumed due to a re-localisation effect upon co-expression. **b** Interactions between putative asparagus secondary cell wall XSC members as reported by Zeng *et al* (2016).

In the asparagus system, various combinations of asparagus GT43 and GT47 IRX orthologues were transiently expressed in *N. benthamiana* leaves; microsomes prepared from these leaves were assayed for xylan xylosyltransferase activity. Maximal activity was seen when *Ao*IRX9, *Ao*IRX14A, and *Ao*IRX10 were co-expressed (Zeng *et al*, 2016). Native-PAGE analysis of microsomal protein appeared to indicate the formation of a high molecular weight complex containing all three proteins (Zeng *et al*, 2016). Furthermore, BiFC experiments on fluorescently tagged versions of these proteins (also expressed in *N. benthamiana* leaves) indicated the presence of direct interactions between *Ao*IRX9 and *Ao*IRX14A, but not between either *Ao*IRX9 and *Ao*IRX10 or *Ao*IRX10 and *Ao*IRX14A (**Figure 3.1b**). BiFC experiments also revealed that all three proteins likely form homodimers. Furthermore, co-expression of all three proteins appeared to be required for their complete transport from the ER to the Golgi in *N. benthamiana* leaves (Zeng *et al*, 2016).

These results suggest that IRX9/IRX9L, IRX14/IRX14L, and IRX10/IRX10L may also form a functional hetero-oligomer in Arabidopsis—although they do not give a coherent picture of how such a complex may be assembled. Interestingly, in both experiments, the GT43 orthologues were observed to form homodimers; in both cases, these proteins were suggested to maintain their homodimeric state within the XSC. However, this idea is at odds with the concurrently published idea that Golgi GTs transition between mutually exclusive homodimeric and heterodimeric states as they travel through the Golgi (Kellokumpu *et al*,

2016). Furthermore, in a previous *Renilla* luciferase complementation assay, homodimerisation was not detected for any of Arabidopsis IRX9, IRX14, or IRX10 proteins (nor were any interactions detected between them) (Lund *et al*, 2015). Hence, confirmation of the oligomeric state of the Arabidopsis GT43 proteins is likely to be of benefit in understanding XSC assembly.

In fact, it appears that many Golgi GTs form homodimers (Hashimoto *et al*, 2010; Harrus *et al*, 2018). Furthermore, GT homodimerisation has been linked with protein localisation and, in some cases, catalytic activity (Young, 2004; Tu & Banfield, 2010). It appears that any part of a GT has the potential to mediate homodimerisation—be it the catalytic, stalk, transmembrane, or cytoplasmic domain (Young, 2004; Kellokumpu *et al*, 2016). However, due to the (perhaps inherent) lack of crystallographic data, the exact mechanism by which the non-globular domains bring about homodimerisation has not been determined.

Prior to the commencement of this project, I established that IRX9/9L and IRX14/14L proteins both possess highly conserved motifs in their predicted transmembrane domains. Although this was previously reported for IRX14 (Jiang *et al*, 2016), the function of these motifs has not been investigated. Furthermore, I noted the resemblance between these motifs and the well characterised GAS_{right} motif, which is frequently found in transmembrane helix (TMH) dimers (see *Section 1.3.2*). Curiously, GAS_{right} motifs, which are well characterised in plasma membrane proteins (Teese & Langosch, 2015), have not been studied before in the context of Golgi GT dimerisation.

In the past, transmembrane oligomerisation has frequently been studied using variants of the TOXCAT technique, which makes use of the homodimeric DNA-binding domain of the ToxR transcriptional activator from *Vibrio cholerœ* (Russ & Engelman, 1999). TOXGREEN is a more recent variant of this technique that facilitates the quantitation step by replacing the canonical chloramphenicol resistance reporter assay with GFP (Armstrong & Senes, 2016). The technique requires the construction of a fusion protein with the ToxR DNA activator at its N-terminus, the MalE maltose binding protein (MBP) at its C-terminus, and a transmembrane helix sequence of interest linking the two (*i.e.* ToxR-TMH-MBP). This protein is expressed in *E. coli* MM39 (a *malE* deletion strain) using a pccGFP vector. Under ideal conditions the ToxR-TMH-MBP fusion protein is typically localised to the inner membrane such that the MBP domain resides in the periplasm (Kolmar *et al*, 1995), thereby permitting MM39 to grow on maltose-supplemented minimal medium (Russ & Engelman, 1999). Dimerisation of the

TMH in the membrane permits the two peripheral ToxR domains to adopt an active homodimeric conformation (Miller *et al*, 1987; DiRita & Mekalanos, 1991). ToxR can then bind to the *ctx* promoter on the pccGFP plasmid in order to activate the transcription of the GFP reporter.

Accordingly, I set out to investigate the transmembrane motifs in these plant GT43s using TOXGREEN. In this chapter, I show that the transmembrane domains of IRX9 and IRX14 likely homo-oligomerise via separate GAS_{right} motifs. I also show that the transmembrane dimerisation of IRX9 is essential for xylan production during secondary cell wall synthesis, and that this could be due to a role of transmembrane dimerisation in proper localisation. These results highlight the importance of glycosyltransferase homodimerisation in function and strengthen our knowledge of how the components of the XSC are assembled.

3.2 Results

3.2.1 Sequence alignments reveal strongly conserved motifs in the transmembrane regions of plants GT43s

To better characterise the transmembrane motifs of IRX9 and IRX14, all GT43 protein sequences from the PlantCAZyme database (Ekstrom *et al*, 2014) were aligned, and a GT43 hidden Markov model (HMM) was constructed using HMMER (Eddy, 2011). This GT43 HMM was used to search for homologues in four additional algal/lower plant genomes (*Klebsormidium nitens*, *Mesotœnium endlicherianum, Anthoceros angustus*, and *Salvinia cucullata*) in order to extend the evolutionary coverage of the analysis. The combined 240 GT43 sequences (including *Hs*GlcAT-I, for rooting) were realigned, and a phylogeny was constructed using FastTree (Price *et al*, 2010). This phylogeny was then used to sort the sequences into two groups: GT43-A (IRX9/9L-related) and GT43-B (IRX14/14L-related). A region containing the predicted transmembrane helix and the flanking 20 amino acids from both sides was extracted from each sequence. New alignments of these extracted regions were then constructed (using a high gap-opening penalty) for the GT43-A and GT43-B groups. Construction of a consensus logo for each, using WebLogo (Crooks *et al*, 2004), revealed a striking pattern of residues within the TMH, with an extremely high level of conservation—in contrast to the surrounding amino acids on either side of the TMH (**Figure 3.2**). These motifs resemble GAS_{right} transmembrane dimerisation motifs, especially in the GT43-A (IRX9) group, wherein the motif constitutes 'FxxGxxxG'. These results suggest that transmembrane sequence motifs are universal in GT43 proteins from throughout the plant kingdom, and that therefore, they likely carry out an important function.

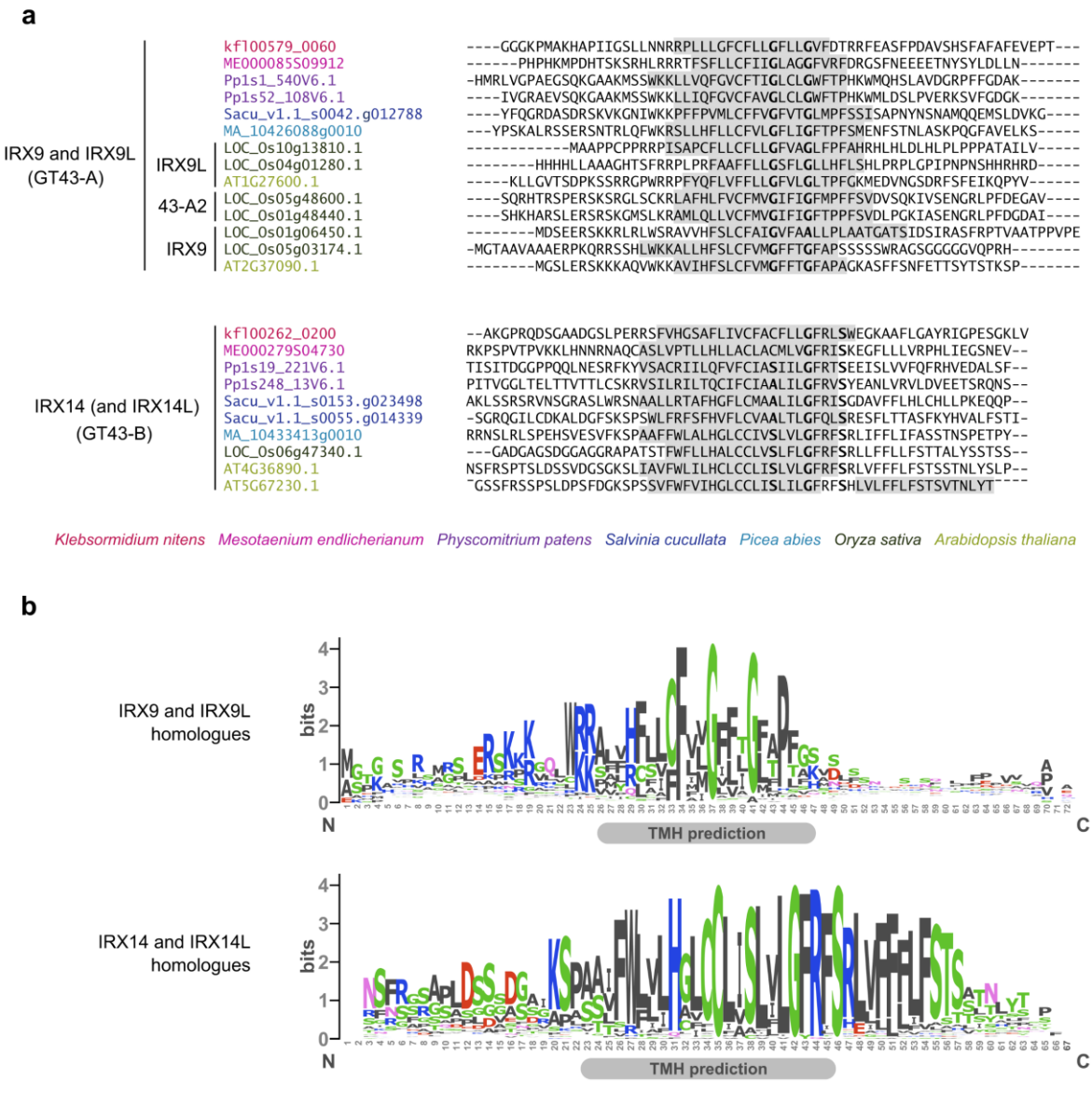


Figure 3.2 The predicted transmembrane domains of plant GT43-family proteins contain conserved GAS_{right} motifs. TMHMM was used to identify potential TMHs in 163 GT43-A (IRX9/9L) sequences and 76 GT43-B (IRX14/14L) sequences. Each TMH-containing sequence was truncated to a region beginning 20 residues upstream of the predicted TMH and 20 residues downstream of its C-terminus before alignment. **a** Aligned TMH regions from a small selection of species. The predicted TMH is highlighted in grey; GAS_{right} motif residues are in bold text. **b** The alignments of the full set of sequences were submitted to WebLogo 3 (http://weblogo.threeplusone.com). The stack height at each site shows the difference between the maximum entropy possible and the entropy of the observed distribution (in bits).

3.2.2 IRX9 and IRX14 transmembrane helices form homo-oligomers when expressed in E. coli

To investigate the hypothesis that these conserved residues might constitute functional GAS_{right} motifs, the propensity for isolated *At*IRX9 and *At*IRX14 transmembrane helices to homodimerise when expressed in *E. coli* was examined using TOXGREEN, a fluorescent reporter-based assay (Armstrong & Senes, 2016). In this technique, truncation of the TMH of interest must be optimised in order to find a suitable fragment than can be properly inserted into the *E. coli* inner membrane. Hence, initially, three IRX9 and four IRX14 truncations were trialled for homodimerisation and correct membrane insertion. Candidate TMH sequences were cloned into an expression plasmid such that they became flanked by sequences encoding the N-terminal ToxR DNA activation domain and the C-terminal maltose binding protein (MBP) domain. The resultant fusion protein was expressed in *E. coli* MM39 (lacking native MBP) alongside a cassette for (dimeric-)ToxR-inducible expression of GFP (**Figure 3.3a,b**). In addition, a ToxR-TMH-MBP fusion containing the TMH of either wild-type or [G83I]-mutant human glycophorin A (GpA and GpA*) was used as a positive or negative control for transmembrane dimerisation, respectively. After growth to stationary phase in liquid LB culture, I measured the GFP fluorescence of three transformants per construct. Of the IRX9/IRX14 constructs, all but one exhibited a large increase in fluorescence relative to the negative control (**Figure 3.3c,d**). However, in contrast to the GpA controls, transformants of all IRX9/IRX14 constructs failed to grow on minimal medium supplemented with maltose, demonstrating that the MBP domain was not properly localised to the periplasmic side of the inner membrane in these cells (**Figure 3.4**). These results indicate that, although these TMH truncations were potentially able to form homodimers, they could not be correctly inserted into the *E. coli* inner membrane, which prevents a valid interpretation of their dimerisation behaviour.

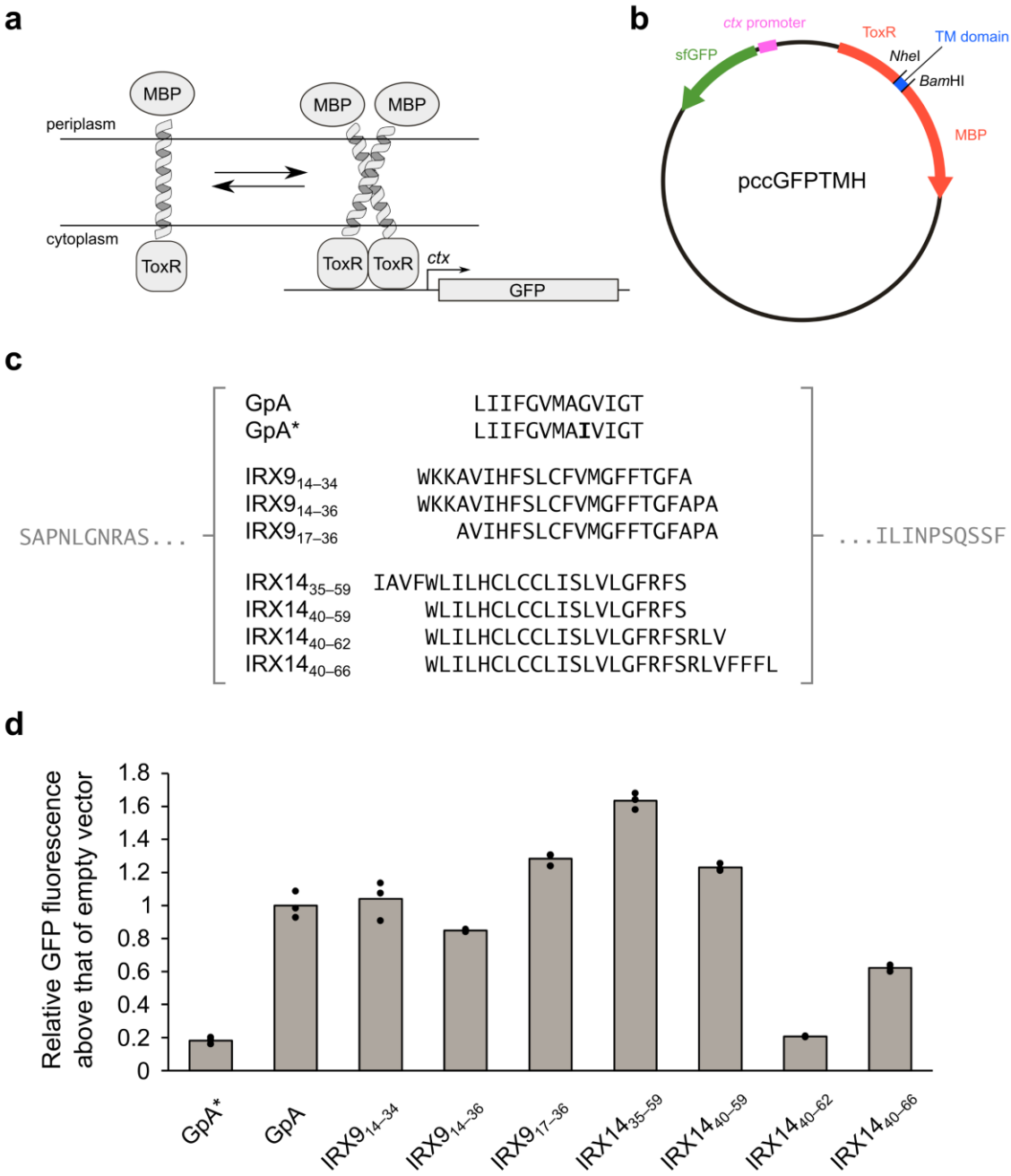


Figure 3.3 Assaying GT43 transmembrane domain dimerisation using TOXGREEN. **a** Schematic of TOXGREEN assay, adapted from Armstrong *et al*, 2016. TMH dimerisation activates the ToxR transcriptional activator and therefore induces expression of GFP from the pccGFPTMH plasmid, provided the N-terminus is correctly localised to the cytoplasmic side of the membrane. **b** Plasmid map of pccGFPTMH, adapted from Armstrong *et al*, 2016. **c** Transmembrane inserts studied. GpA and GpA* are positive and negative controls, respectively. **d** GFP fluorescence of MM39 stationary-phase cultures tranformed with the various pccGFPTMH constructs. Three independent transformants per construct were cultured in LB medium overnight before measurement of OD_{590} and fluorescence at excitation/emission wavelengths of 485/520 nm. Fluorescence readings were normalised to OD_{590} readings before subtraction of the mean fluorescence for empty-vector transformants. Data for individual transformants are shown as black dots; bars show the group mean.

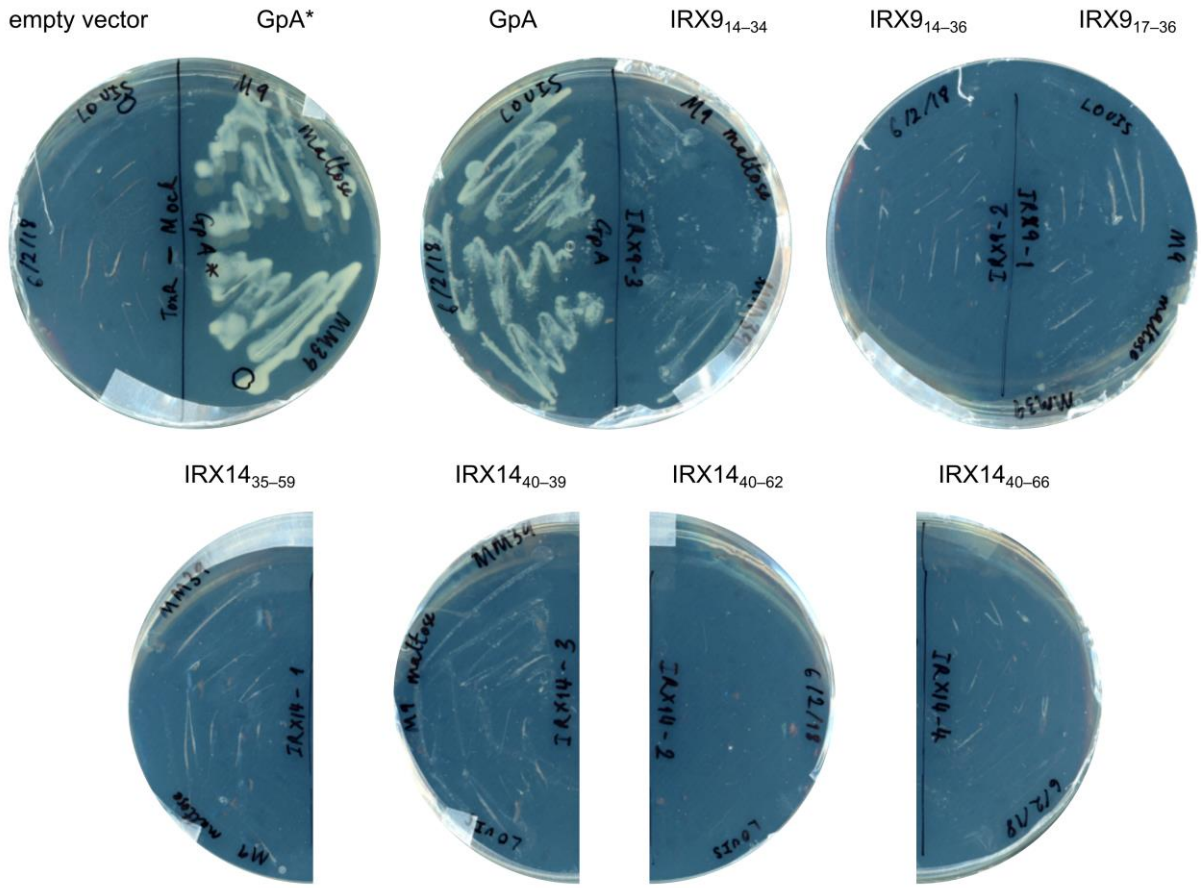


Figure 3.4 Ability of MM39 pccGFPTMH transformants to take up maltose. Transformants (one per construct) were streaked onto M9 minimal medium supplemented with 0.4 % maltose and incubated for 72 h at 37 °C.

To find alternative truncations that *could* be inserted into the *E. coli* inner membrane, a further three IRX9 and three IRX14 TMH truncations were tested for dimerisation and proper localisation. Since the first set of constructs incorporated rather long TMH regions, I prepared shorter TMH inserts for this second round of experiments (considering that downstream hydrophobic residues in the TMH-MBP linker might also be included in the biological TMH). Accordingly, I transformed MM39 cells with the new set of constructs and measured their fluorescence and ability to take up maltose. Although the expression of all six new constructs brought about an increase in GFP fluorescence relative to the GpA* negative control (**Figure 3.5**), the results of the maltose growth assay suggested a range of capacities for membrane insertion (**Figure 3.6a**). Amongst these, $IRX9_{21–34}$ and $IRX9_{21–37}$ had the best apparent membrane insertion, though neither appeared to facilitate maltose uptake quite as well as the two glycophorin A controls. Furthermore, although fluorescence measurements were normalised to cell density, I considered that the level of observed fluorescence could be affected by the abundance of the individual ToxR-TMH-MBP protein. Therefore, I attempted to quantify the level of expression by Western blotting. Using an anti-MBP antibody, I was able to detect two bands in total-protein extracts from ToxR-TMH-MBP-expressing cultures that were not present in the empty vector control—attributed to ToxR-TMH-MBP and free MBP, respectively. Interestingly, despite producing the highest level of GFP fluorescence amongst the IRX9/IRX14 constructs, the $IRX9_{21–34}$ construct appeared to exhibit the joint-lowest expression of full-length protein (**Figure 3.6b,c**). Therefore, when accounting for protein abundance, $IRX9_{21–34}$ still appeared to exhibit both the highest fluorescence and one of the highest maltose uptakes of any construct. Consequently, this particular construct was selected as the best candidate for further replication and point mutation experiments.

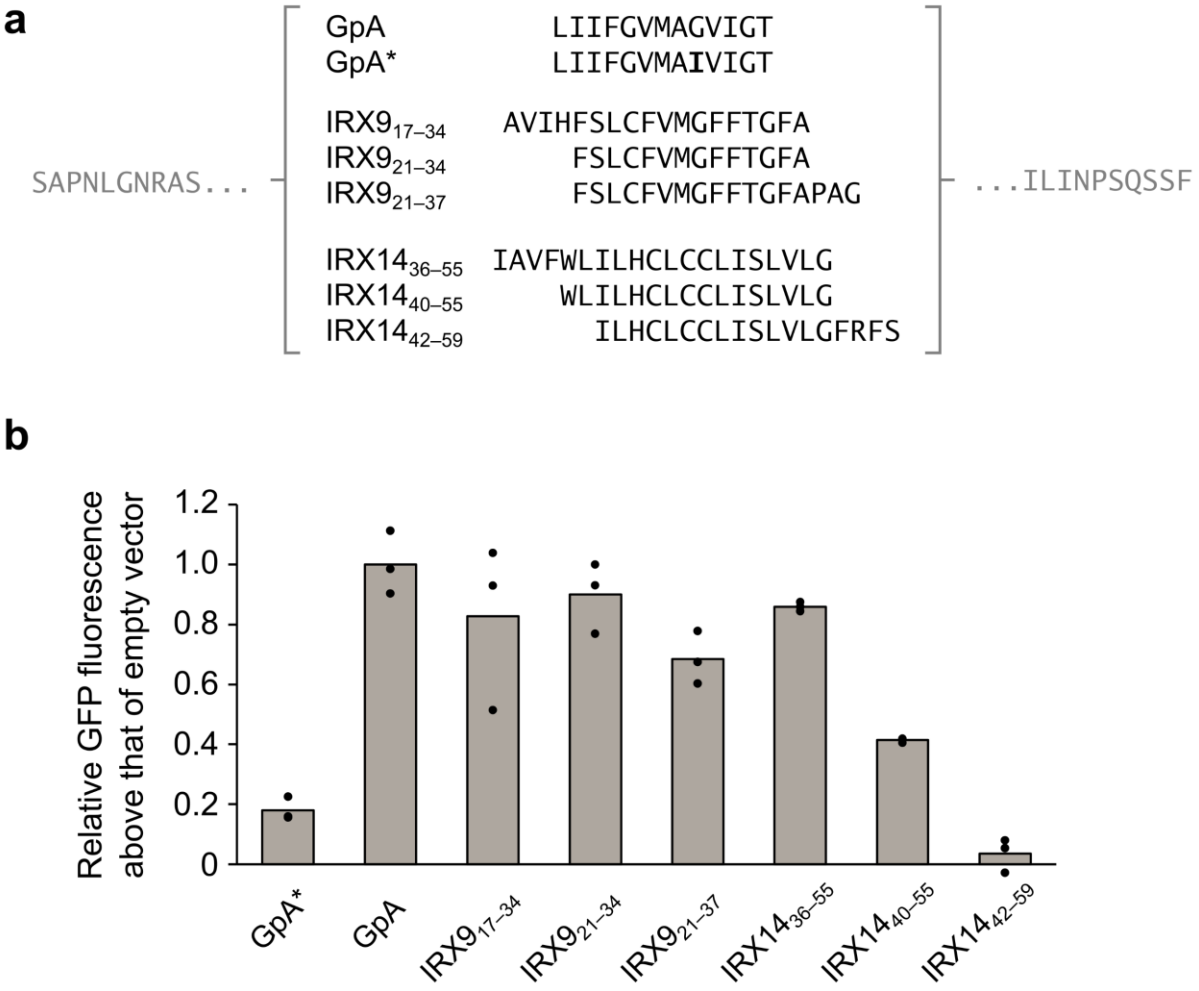
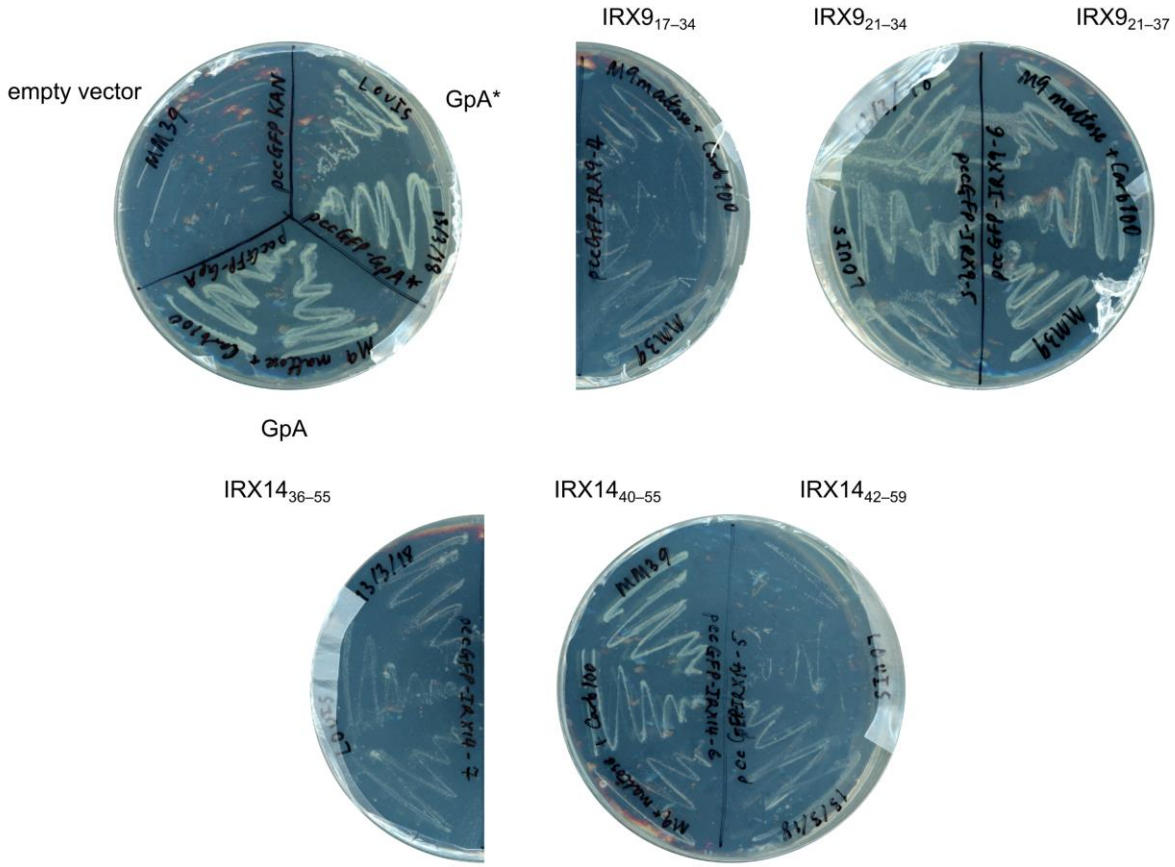
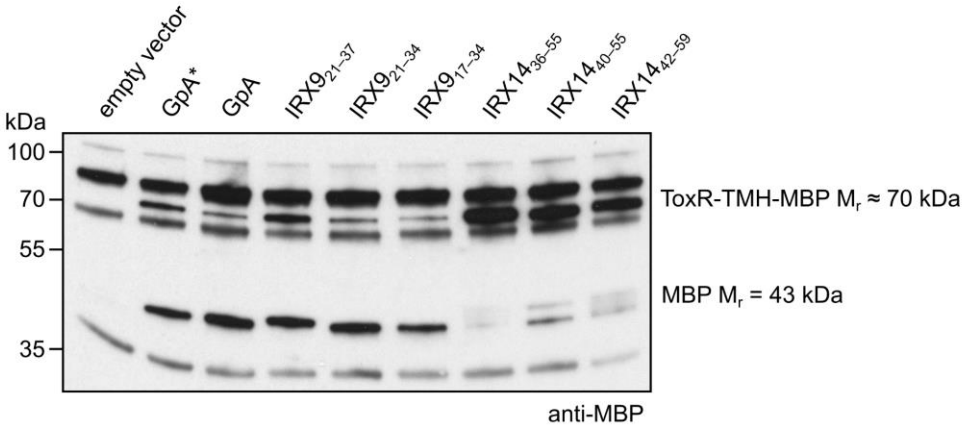


Figure 3.5 Second round of TOXGREEN experiments, using shorter TMH inserts. a Transmembrane inserts studied. GpA and GpA* are positive and negative controls, respectively. **b** GFP fluorescence of MM39 stationary-phase cultures tranformed with the various pccGFPTMH constructs. Three independent transformants per construct were cultured in LB medium overnight before measurement of OD_{590} and fluorescence at excitation/emission wavelengths of 485/520 nm. Fluorescence readings were normalised to OD_{590} readings before subtraction of the mean fluorescence for empty-vector transformants. Data for individual transformants are shown as black dots; bars show the group mean.

a


empty vector
GpA*
GpA
IRX9$_{17–34}$
IRX9$_{21–34}$
IRX9$_{21–37}$
IRX14$_{36–55}$
IRX14$_{40–55}$
IRX14$_{42–59}$

b


empty vector
GpA*
GpA
IRX9$_{21–37}$
IRX9$_{21–34}$
IRX9$_{17–34}$
IRX14$_{36–55}$
IRX14$_{40–55}$
IRX14$_{42–59}$
kDa
100
70
55
35
ToxR-TMH-MBP M_r ≈ 70 kDa
MBP M_r = 43 kDa
anti-MBP

c

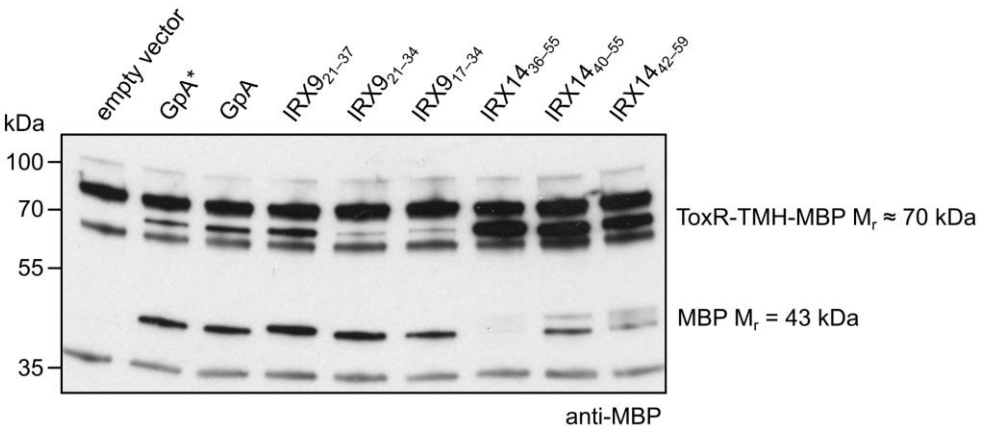

empty vector
GpA*
GpA
IRX9$_{21–37}$
IRX9$_{21–34}$
IRX9$_{17–34}$
IRX14$_{36–55}$
IRX14$_{40–55}$
IRX14$_{42–59}$
kDa
100
70
55
35
ToxR-TMH-MBP M_r ≈ 70 kDa
MBP M_r = 43 kDa
anti-MBP

Figure 3.6 Second round of TOXGREEN experiments: maltose uptake of transformants and expression of the ToxR-TMH-MBP protein. a MM39 pccGFPTMH transformants (one per construct) were streaked onto M9 minimal medium supplemented with 0.4 % maltose and incubated for 72 h at 37 °C. **b** Total protein was extracted from transformants (one per construct) and separated by SDS-PAGE. The presence of the C-terminal MBP domain was then probed by Western blotting. **c** Same as **b** but using a different transformant for each construct, demonstrating the lack of substantial variation in expression between biological replicates.

3.2.3 IRX9 TMH oligomerisation is dependent on its GAS_{right} motif

To investigate the requirement of the FxxGxxxG motif for the dimerisation of the ToxR-$IRX9_{21–34}$-MBP protein in the *E. coli* inner membrane, I made a new version of the $IRX9_{21–34}$ construct in which the first glycine of the GAS_{right} motif (Gly28) was mutated to isoleucine ($IRX9_{21–34}$[G28I], or $IRX9_{21–34}$* for short). Unlike glycine, isoleucine—a relatively common amino acid in plant Golgi transmembrane helices (Parsons *et al*, 2019)—possesses a bulky sidechain, and, when placed at such a position, blocks the formation of GAS_{right} dimers due to steric clashes with the main chain of the opposing helix (Lemmon *et al*, 1992; Russ & Engelman, 1999). I also designed additional constructs containing one further IRX9 truncation and two further IRX14 truncations, with the hope that the resultant proteins might exhibit improved membrane integration compared with the previous round of constructs. As before, I transformed all new constructs into *E. coli* MM39 and measured the fluorescence, ability to take up maltose, and protein expression. As expected, the new $IRX9_{20–34}$, $IRX14_{41–55}$, and $IRX14_{42–55}$ constructs, as well as the previously characterised $IRX9_{21–34}$ construct, gave rise to substantial GFP fluorescence (**Figure 3.7**); however, none of the relevant transformants could take up maltose as well as the GpA controls (**Figure 3.8**). In contrast, transformants of the $IRX9_{21–34}$* construct did not exhibit any increased GFP fluorescence compared to the negative control. Interestingly, these transformants also appeared better able to take up maltose. These results strongly suggest that the FxxGxxxG motif seen in the IRX9 TMH constitutes a canonical GAS_{right} motif capable of facilitating homodimerisation; however, a better understanding of why the unmutated $IRX9_{21–34}$ fusion protein does not appear to integrate well into the *E. coli* inner membrane may be required in order to fully interpret these results.

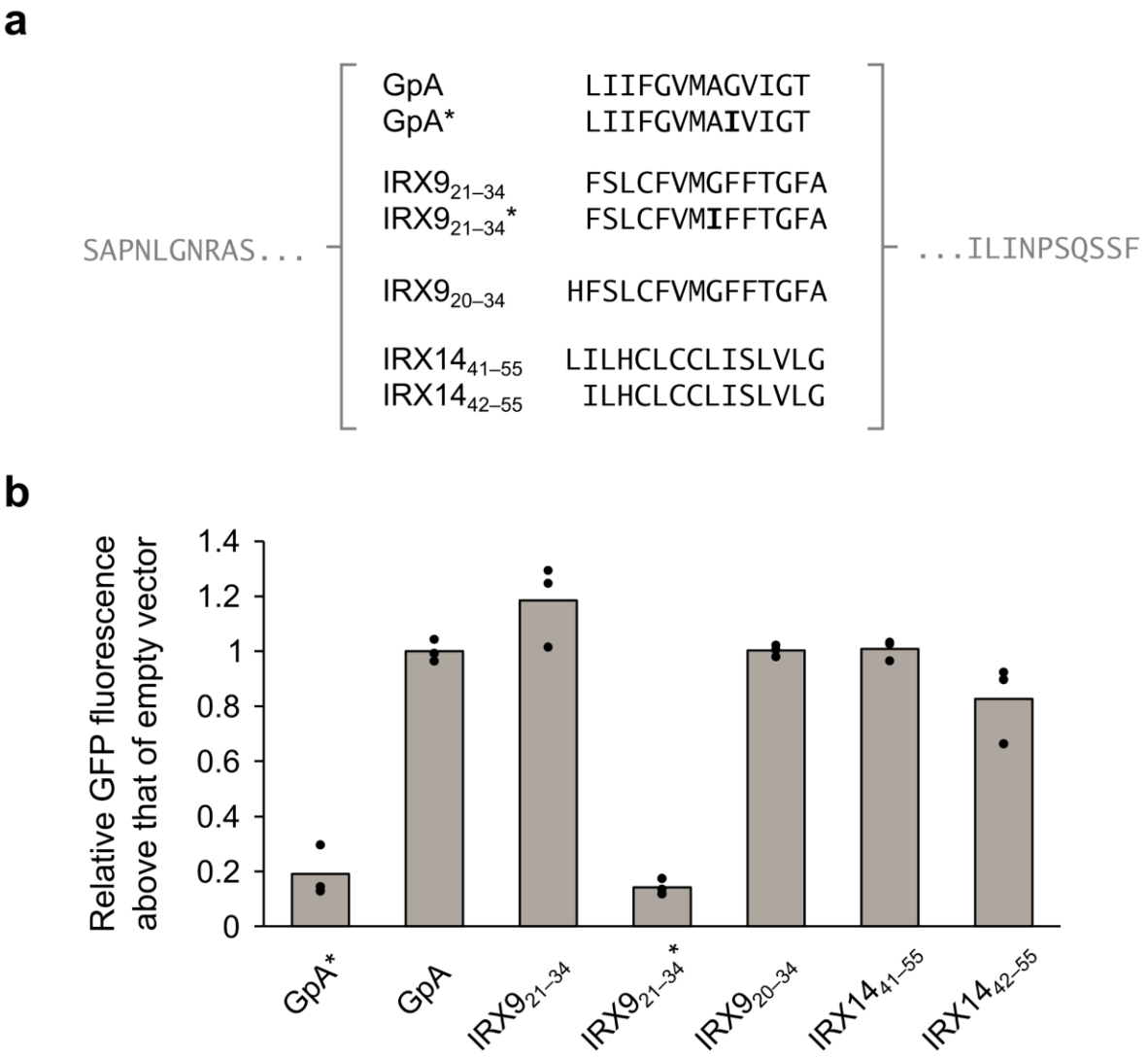


Figure 3.7 Third round of TOXGREEN experiments: mutation of Gly28 to isoleucine abrogates TMH homo-dimerisation. a Transmembrane inserts studied. GpA and GpA* are positive and negative controls, respectively. **b** GFP fluorescence of MM39 stationary-phase cultures tranformed with the various pccGFPTMH constructs. Three independent transformants per construct were cultured in LB medium overnight before measurement of OD_{590} and fluorescence at excitation/emission wavelengths of 485/520 nm. Fluorescence readings were normalised to OD_{590} readings before subtraction of the mean fluorescence for empty-vector transformants. Data for individual transformants are shown as black dots; bars show the group mean.

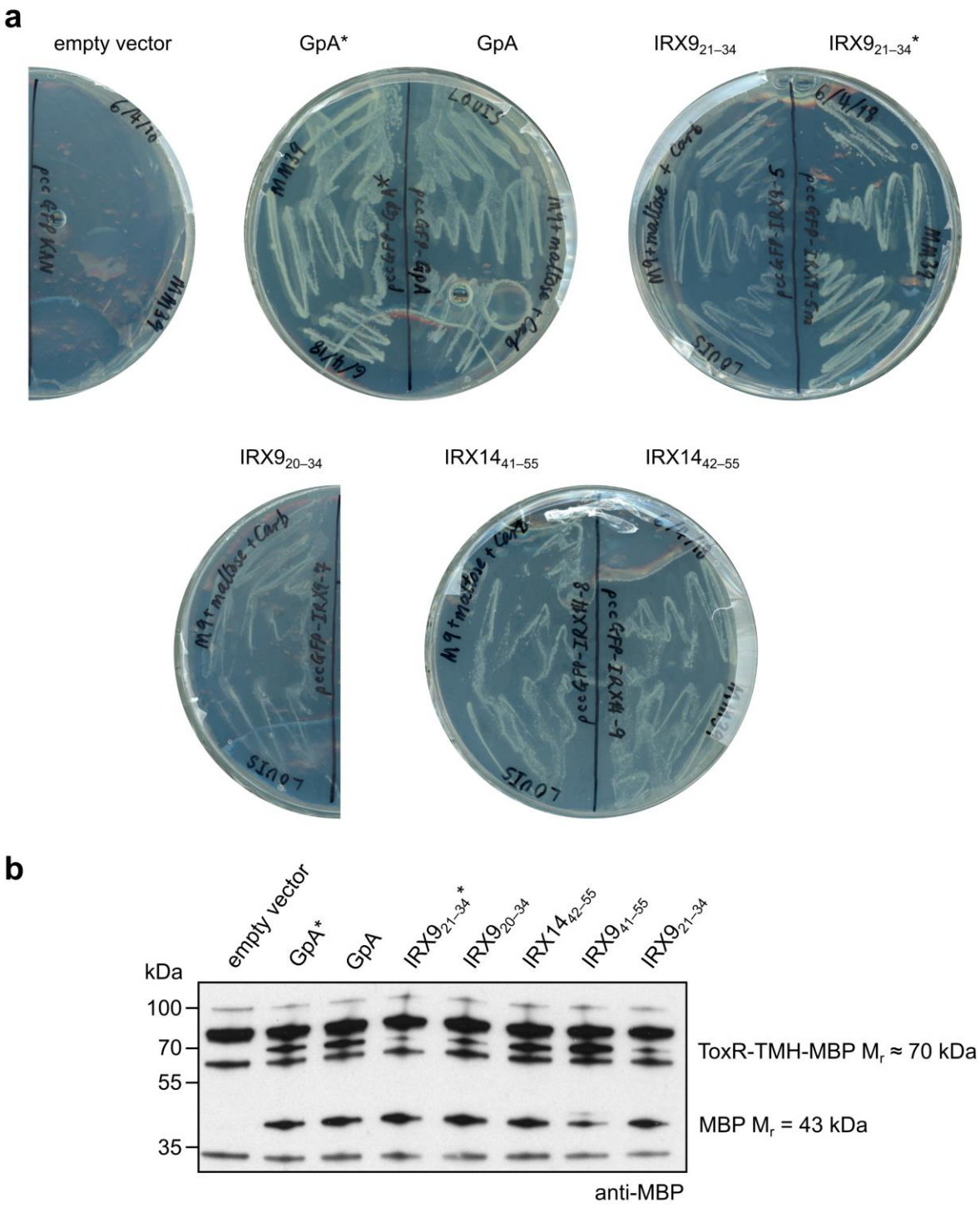


Figure 3.8 Third round of TOXGREEN experiments: maltose uptake of transformants and expression of the ToxR-TMH-MBP protein. a MM39 pccGFPTMH transformants (one per construct) were streaked onto M9 minimal medium supplemented with 0.4 % maltose and incubated for 72 h at 37 °C. **b** Total protein was extracted from transformants (one per construct) and separated by SDS-PAGE. The presence of the C-terminal MBP domain was then probed by Western blotting.

3.2.4 *Disruption of the GAS$_{right}$ motif prevents IRX9 function* in planta

The above results concern the behaviour of truncated IRX9 in an *E. coli* lipid bilayer; however, they do not demonstrate the importance of the FxxGxxxG motif *in planta*. Furthermore, in addition to the GAS$_{right}$ motif, the IRX9 TMH contains a conserved cysteine (Cys24)—since many animal GTs have been reported to form intermolecular disulphide bridges in their transmembrane regions (Tu & Banfield, 2010), it is possible that this cysteine is also involved in TMH homodimerisation. Hence, I wanted to investigate the function of these residues in Arabidopsis. Accordingly, to determine whether the IRX9 FxxGxxxG motif, or the cysteine preceding it, play an important role in the function of full-length IRX9, *irx9* mutant Arabidopsis plants were transformed with an expression cassette harbouring either wild-type IRX9, IRX9[G28I], or IRX9[C24S] downstream of the promoter of IRX14 (all three were C-terminally tagged with c-Myc). In order to determine whether the expression of these proteins could complement the *irx9* phenotype, and thereby assess their ability to function in the place of endogenous IRX9, I observed the overall phenotype of T_2-generation transgenic plants after six weeks of growth. All plants were homozygous at this stage except for line 3 of the *irx9* pIRX14::IRX9 plants, for which, for unknown reasons, no homozygous offspring could be obtained. The *irx9* plants expressing IRX9 or IRX9[C24S] appeared essentially identical to wild-type plants, suggesting that both proteins are able to complement the mutant phenotype when expressed under the IRX14 promoter (**Figure 3.9**). Interestingly, however, plants expressing IRX9[G28I] exhibited phenotypes much closer to that of *irx9*, suggesting either that IRX9[G28I] is not expressed to the same level, or that this protein is less functional than IRX9 or IRX9[C24S].

The stunted growth of the *irx9* mutant appears to be caused by the collapse of vessels and fibres in the stem, in turn due to a lack of xylan in the secondary cell walls of these cells (Brown *et al*, 2005, 2007). However, stunted-growth phenotypes might conceivably be caused by other factors apart from a lack of xylan. Therefore, to determine whether the observed growth phenotypes could be due to changes in the amount of secondary cell wall xylan, I wanted to characterise bottom-stem sections of homozygous Arabidopsis lines by monosaccharide analysis. Accordingly, I prepared alcohol-insoluble residue (AIR) from the bottom fifth of each stem, subjected the material to total carbohydrate hydrolysis with trifluoroacetic acid, and quantified the monosaccharides released. Both *irx9* and *irx9* pIRX14::IRX9[G28I] plants exhibited a clear reduction in overall xylose content compared with wild-type plants, whereas *irx9* pIRX14::IRX9, and *irx9* pIRX14::IRX9[C24S] plants exhibited similar levels to wild-type

(**Figure 3.10**). These results confirm that the phenotype of the *irx9* pIRX14::IRX9[G28I] plants is likely due to a reduction in xylan synthesis compared with wild-type plants—supporting the notion that the FxxGxxxG motif is essential for the role of IRX9 in xylan synthesis.

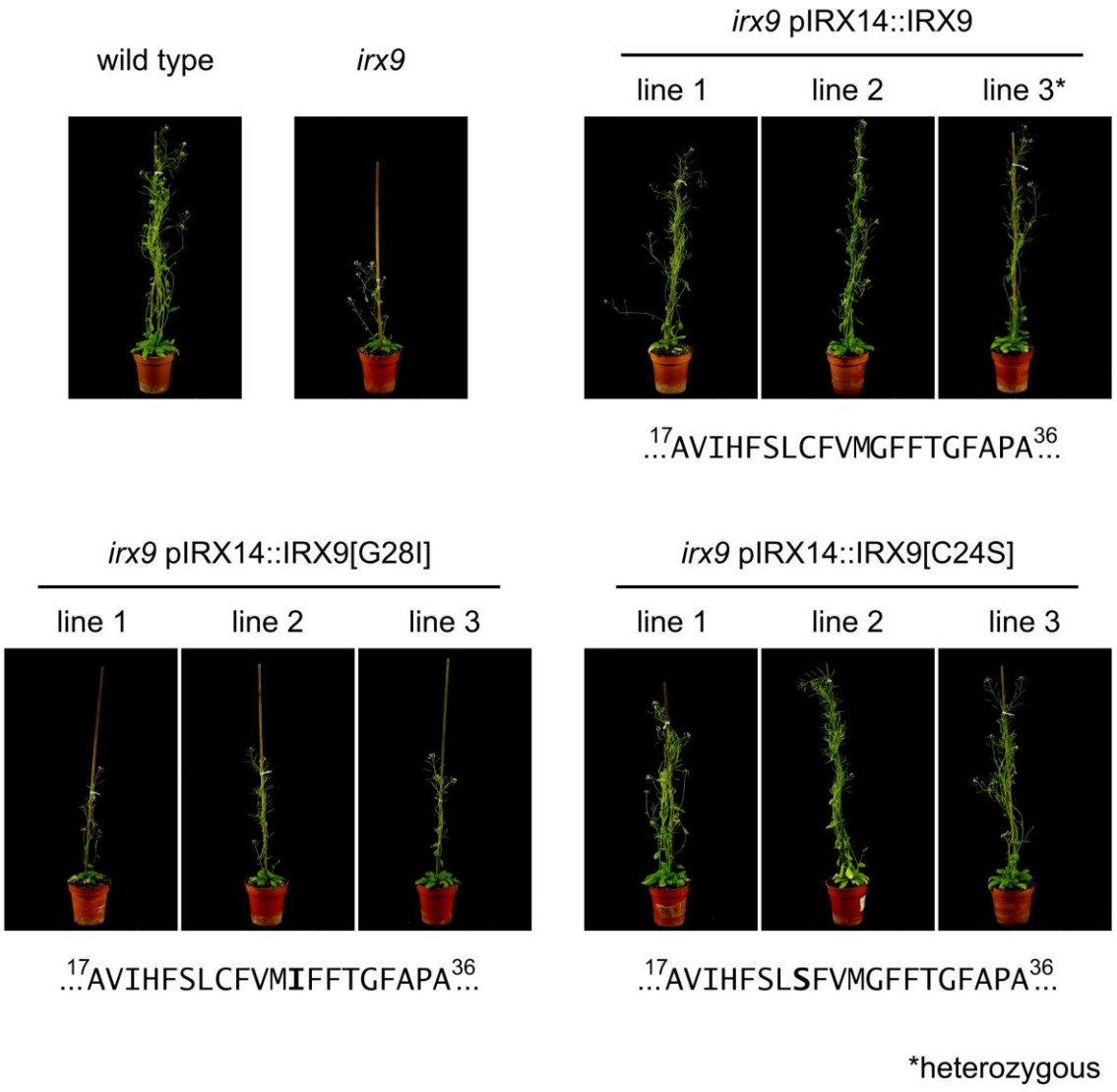


Figure 3.9 Expression of IRX9 or IRX9[C24S], but not IRX[G28I], is able to fully rescue the *irx9* phenotype. Arabidopsis *irx9* plants were transformed by floral dip to introduce constructs for expression of IRX9, IRX9[G28I], or IRX9[C24S] under the control of the IRX14 promoter. Plants were bred to the T_2 generation, and grown for 6 weeks before photographing representative individuals. Homozygosity was later assessed by seed genotyping; however, no homozygous plants were identified for line 3 of *irx9* pIRX14::IRX9, and further breeding to produce homozygotes from the hemizygotes was unsuccessful.

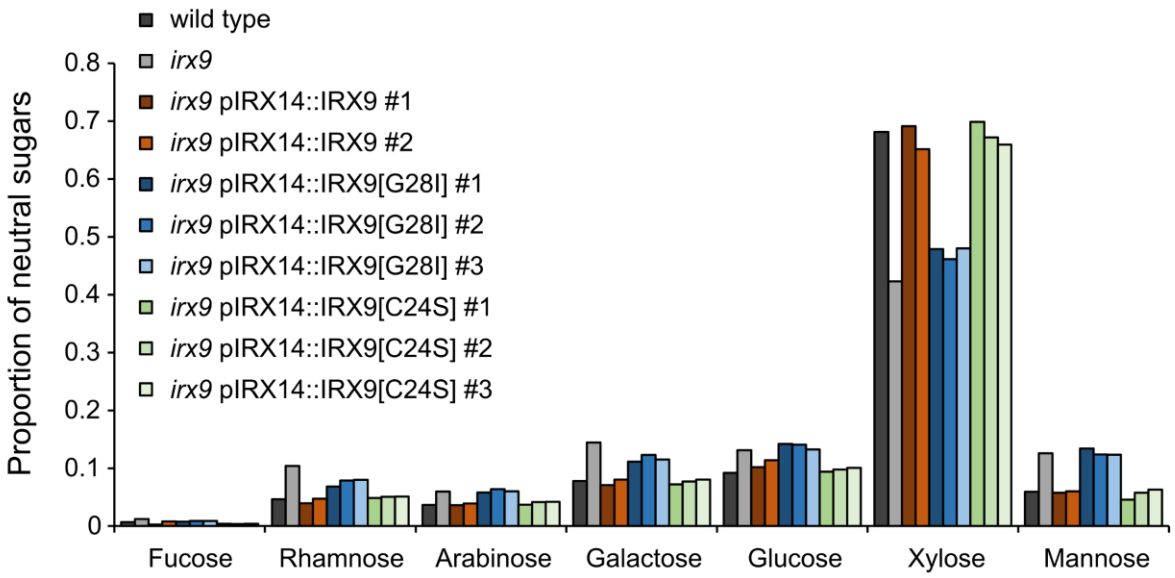


Figure 3.10 Monosaccharide analysis reveals that the expression of IRX9[G28I] does not restore *irx9* bottom-stem xylose content to wild-type levels. Alcohol-insoluble residue (AIR) was prepared from the bottom stems of wild-type, untransformed *irx9*, and transgenic *irx9* plants before acid hydrolysis using 2 M trifluoroacetic acid. For each transgenic line, all homozygous plants were pooled together before preparation of AIR. Monosaccharides were identified and quantified by HPAEC-PAD using standards of known concentration. Bars show the quantity of each sugar in each genotype as a proportion of total detected neutral sugars. Due to equipment failure, only single measurements could be made for each genotype.

3.2.5 Disruption of the GAS$_{right}$ motif in IRX9 alters its subcellular localisation

Because IRX9 has been proposed to fulfil a structural role in the XSC (Ren *et al*, 2014), the possibility that its oligomerisation might be important for xylan synthesis is not wholly unanticipated. However, why the apparent homodimerisation of its transmembrane domain, specifically, should be so essential for its function is not as obvious. Nevertheless, because the transmembrane domain attaches the globular domain to the Golgi membrane, and also because perturbation of homodimerisation has been previously shown to affect the trafficking of some GTs (Tu & Banfield, 2010), I hypothesised that the homodimerisation of the IRX9 TMH may have a role in its correct localisation. Hence, I set out to study the subcellular localisation of IRX9[G28I].

The transient over-expression of fluorescently tagged proteins in *Nicotiana benthamiana* leaves has frequently been employed to study the localisation of plant GTs (Amos & Mohnen,

2019). However, the study of IRX9 localisation in particular is potentially more complicated because its *Asparagus officinalis* orthologue, *Ao*IRX9, appears to require the co-expression of IRX10 and IRX14 orthologues, *Ao*IRX10 and *Ao*IRX14A, for its transport from the ER to the Golgi in the *N. benthamiana* system (Zeng *et al*, 2016). Therefore, before designing point mutation experiments, I wanted to ascertain whether Arabidopsis IRX9 similarly requires the co-expression of IRX10 and IRX14 for Golgi localisation in *N. benthamiana* leaves. To that end, I prepared two constructs for transient expression. The first contained IRX9 C-terminally tagged with GFP (IRX9-GFP) as well as a *cis*-Golgi marker (the N-terminus of mannosidase I from *Glycine max*) C-terminally tagged with the red fluorescent protein mCherry (ManI-mCherry) (plasmid EC62254 in *Section 2.3.1*). The second construct, on the other hand, contained IRX9-GFP and ManI-mCherry in addition to IRX10-FLAG and IRX14-HA (plasmid EC62252). I transformed *N. benthamiana* leaves with each construct by agroinfiltration. Three days later, epidermal cells in these leaves were examined using confocal microscopy (for this experiment, the microscope was operated by Dr Henry Temple). Fluorescence corresponding to GFP and mCherry was visible in both sets of leaves. In leaves expressing the first construct (IRX9-GFP and ManI-mCherry), the mCherry fluorescence was localised in a punctate pattern consistent with the expected localisation of plant Golgi bodies; however, the GFP fluorescence instead appeared predominantly in a reticulate pattern that, for the most part, did not overlap substantially with the mCherry signal (**Figure 3.11a**). This pattern of GFP fluorescence was interpreted to represent ER localisation, though some punctate spots were also visible. In contrast, in leaves expressing the second construct (IRX9-GFP, IRX10-FLAG, IRX14-HA, and ManI-mCherrry), both GFP and mCherry fluorescence fully co-localised in the punctate Golgi pattern. To characterise this pattern better, I examined the spots of signal at higher magnification. Indeed, whereas mCherry fluorescence was seen at the centre of each Golgi stack, the GFP fluorescence originating from IRX9-GFP was seen only at the periphery of the stack, forming a ring-like shape (**Figure 3.11b**)—exactly as previously described for native Arabidopsis stem (Meents *et al*, 2019). These results suggest that co-expression of Arabidopsis IRX10 and IRX14 are required for the transport of Arabidopsis IRX9 from the ER to the Golgi in *N. benthamiana* leaves.

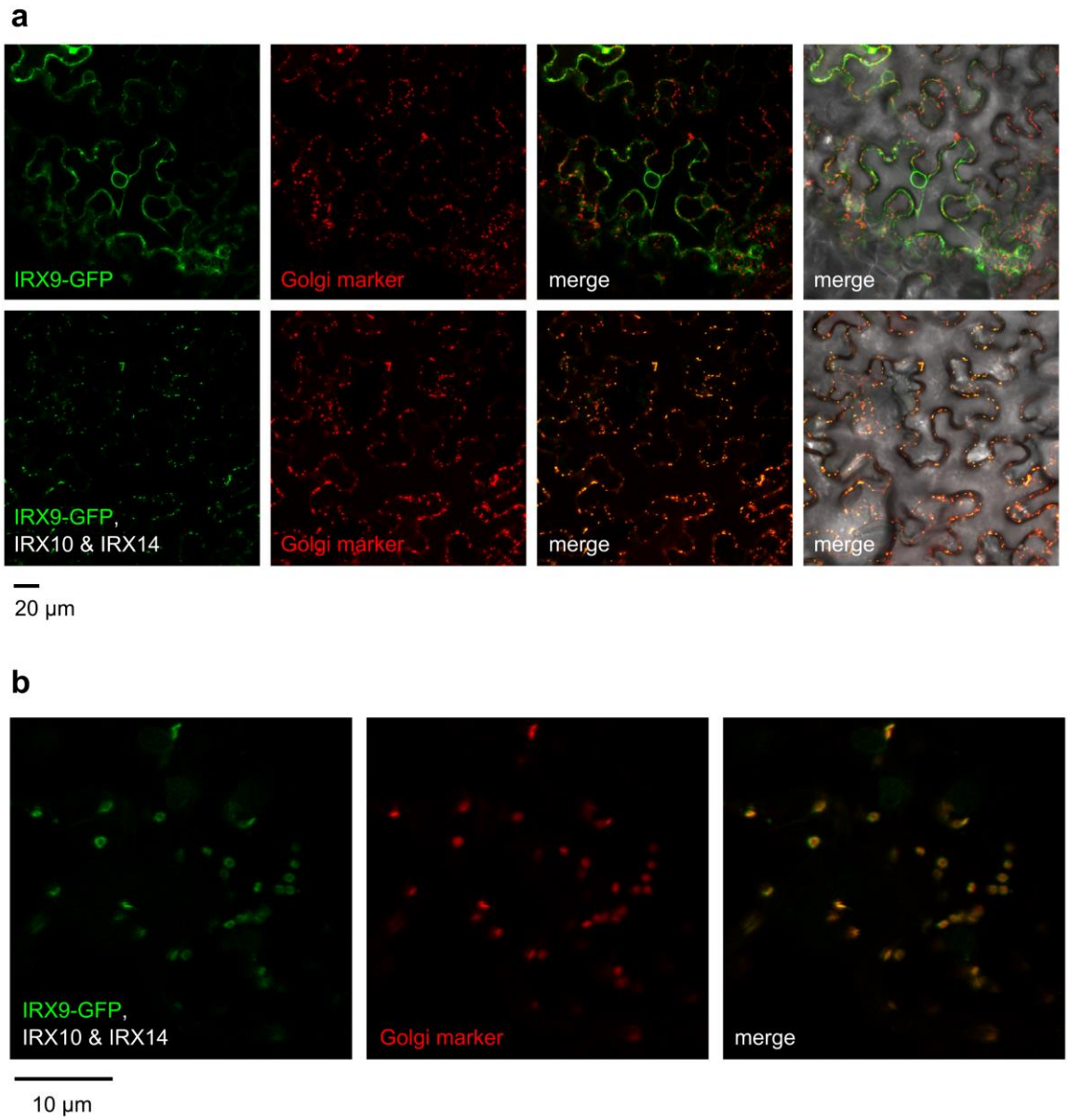


Figure 3.11 When transiently over-expressed in *Nicotiana benthamiana,* Arabidopsis IRX9 requires the co-expression of IRX10 and IRX14 for localisation to the Golgi. *N. benthamiana* leaves were transformed by agroinfiltration; leaf epidermis was examined by confocal microscopy three days post infiltration. **a** Top row: leaves were transformed with a construct for expression of IRX9-GFP and the *cis*-Golgi marker ManI-mCherry. Bottom row: leaves were transformed with a construct for expression of IRX9-GFP, IRX10-FLAG, IRX14-HA, and the *cis*-Golgi marker ManI-mCherry. Left to right: GFP fluorescence, mCherry fluorescence, overlay of the two signals, and overlay of the two signals combined with bright-field view. **b** Higher magnification of fluorescence pattern resulting from expression of the four-gene construct.

Consequently, I created expression constructs containing IRX10-FLAG, IRX14-HA, and one of IRX9-GFP, IRX9[G28I]-GFP, or IRX9[C24S]-GFP (plasmids EC62521–3). Due to technical difficulties in construct assembly, it was necessary to create a separate expression construct for ManI-mCherry in this experiment (plasmid EC62507). *N. benthamiana* leaves were then co-transformed with both the Golgi marker construct and one of the three IRX9/10/14 constructs. After three days, I examined epidermal cells in these leaves using confocal microscopy. I located cells that exhibited both GFP and mCherry fluorescence. As previously, in leaves expressing wild-type IRX9-GFP, IRX10-FLAG, IRX14-HA, and ManI-mCherry, both GFP and mCherry fluorescence was co-localised in a punctate pattern consistent with the expected localisation of Golgi stacks (**Figure 3.12**). GFP fluorescence was more difficult to locate in leaves transformed with the IRX9[C24S] construct; nevertheless, where present, the signal appeared to broadly co-localise with that of mCherry signal, though, due to its low intensity, it was not clear whether the additional non-co-localised signal emanated from IRX9[C24S]-GFP or from background sources. In contrast, leaves expressing the IRX9[G28I] construct exhibited very few cells with a punctate pattern of GFP fluorescence. In the small number of cells in which strong GFP fluorescence could be observed, both GFP and mCherry signals were apparently co-localised in a large aggregate. The extent to which this aggregate could represent an artefact of the expression system is unknown. Nevertheless, these results suggest that mutation of Gly28 to isoleucine in IRX9 prevents its proper localisation, and that the overexpression of this mutated protein may even result in endomembrane system remodelling in *N. benthamiana* epidermal cells.

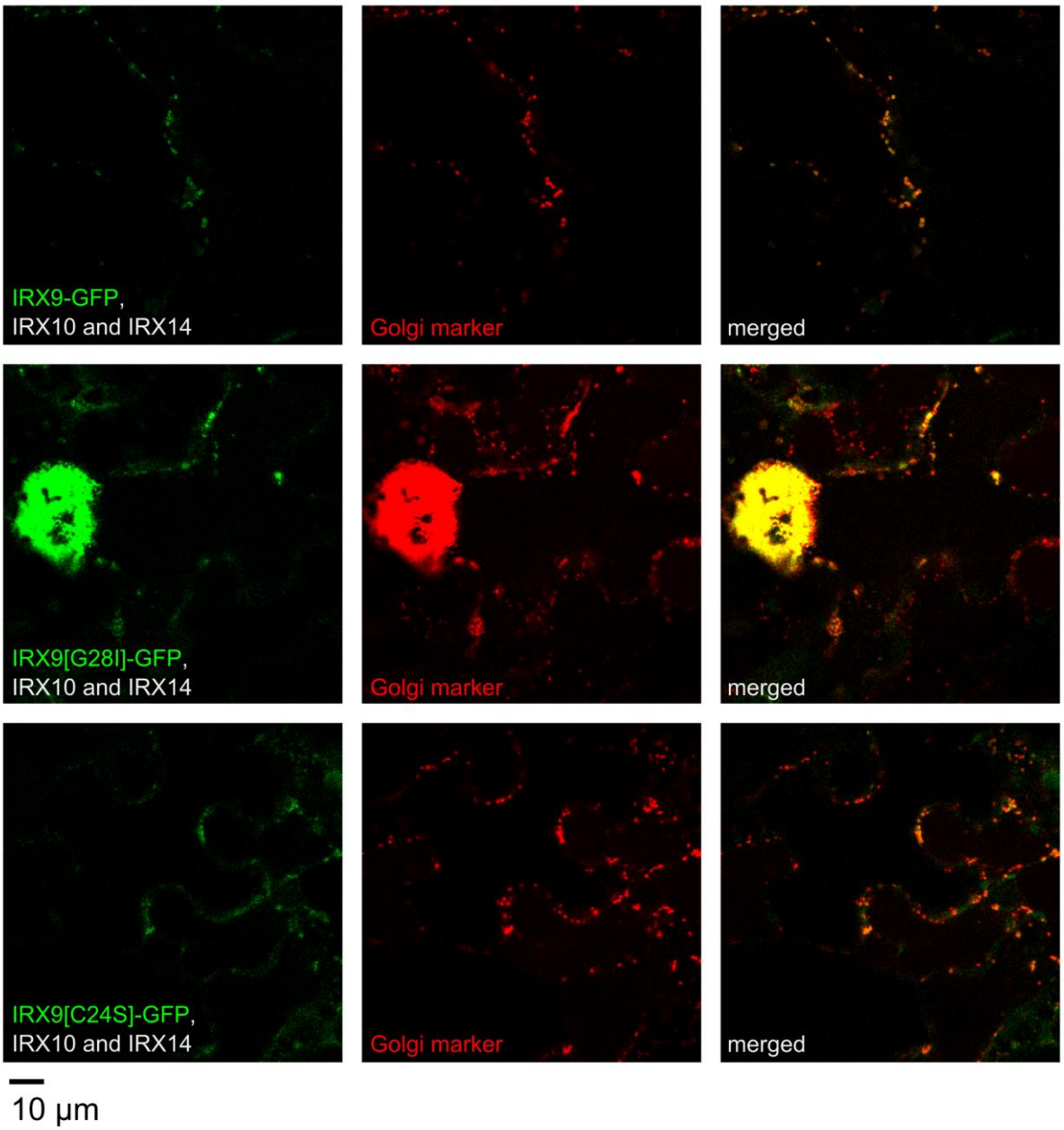


Figure 3.12 When over-expressed in *N. benthamiana*, IRX9[G28I] fails to localise in a normal fashion. *N. benthamiana* leaves were transformed by agroinfiltration; leaf epidermis was examined by confocal microscopy three days post infiltration. Leaves were co-transformed with a construct for expression of ManI-mCherry and a construct for expression of IRX10-FLAG, IRX14-HA, and one of IRX9-GFP (top row), IRX9[G28I]-GFP (middle row), or IRX9[C24S]-GFP (bottom row). Left to right: GFP fluorescence, mCherry fluorescence, and overlay of the two signals. The localisation of IRX9 and mutants is representative of two biological replicates, each examining two plants per construct.

3.2.6 Comparisons between the sequences of IRX9, IRX14, and human GT43 enzymes suggest that the globular domains of IRX9 and IRX14 also form homodimers

Crystallographic structures of human GT43s demonstrate that their catalytic domains form homodimers (Pedersen *et al*, 2000, 2002; Kakuda *et al*, 2004; Shiba *et al*, 2006). Because my TOXGREEN data suggested that the transmembrane domains of IRX9 and IRX14 homodimerise, I hypothesised that their globular ('catalytic') domains might also form homodimeric interactions. To investigate this hypothesis, I wanted to determine whether IRX9 and IRX14 share any of the features involved in homodimerisation in the human GT43 enzymes.

To determine the residues most important for the homodimerisation of human GT43s, the structures and sequences of human GlcAT-I, GlcAT-P, and GlcAT-S were submitted to the protein interface detection server PPCheck (Sukhwal & Sowdhamini, 2013). The results were similar for all three human GT43 structures, with the predicted interaction 'hotspot' residues in GlcAT-I being Arg95, Gln98, Leu224, Arg225, and Val335 (**Figure 3.13a**). I examined the previously published structure of GlcAT-I (Pedersen *et al*, 2000, 2002) to understand how these residues contribute to dimerisation. I observed that these five residues constituted two hotspot areas, one involving the solvent-exposed C-terminus, and the other at the centre of the homodimerisation interface. In the former, Leu224 and Arg225 (from hypervariable loop 2) make interactions with the C-terminal residues in the opposing protomer, including Val335 (**Figure 3.13b**). These interactions appear to be mainly van-der-Waals in nature. The second hotspot, however, involves Arg95 and Gln98 (from the α1 helix), which make symmetrical sidechain-specific interactions with various residues from the other chain. For instance, Gln98 appears to form a hydrogen bond with Thr99 in the opposing protomer; this hydrogen-bonding pair is also conserved in GlcAT-P and GlcAT-S (albeit with Asn replacing Gln). Arg95, on the other hand, forms a symmetrical 'arginine pair' stacking interaction with Arg95 in the opposing protomer—an interaction most likely stabilised by the anionic sidechain of Glu92 (these residues are present in all three enzymes). In addition to these hotspot residues, PPCheck identified a potential intermolecular salt bridge between Arg310 and Glu312 (both residues are situated on the loop connecting β7' to the final α-helix). This interaction is particularly interesting, as Arg310 is also thought to interact with the phosphate groups of the nucleotide sugar (Pedersen *et al*, 2000). Hence, dimerisation could indirectly affect binding of substrates in the active site.

a

```
                    β1              α1              β2              α2          β3
                    --EEEEEEEE----HHHHHHHHHHHHHH----EEEEEEE-----HHHHHHHHHHHH-EEE
HsGlcAT-I           LPTIYVVTPTYARLVQKAELVRLSQTLSLVPRLHWLLVEDAEGPTPLVSGLLAASGLLFT  (133)
HsGlcAT-P           LPTIHVVTPTYSRPVQKAELTRMANTLLHVPNLHWLVVEDAPRRTPLTARLLRDTGLNYT  (142)
HsGlcAT-S           LPTIYAITPTYSRPVQKAELTRLANTFRQVAQLHWILVEDAAARSELVSRFLARAGLPST  (138)
                    ****:.:****:* ******.*:::*:  *..***::****   : *.: :*  :**  *
                                                α3                      β4
                    EEE---------------------HHHHHHHHHHHH----EEEE----------EEEE
HsGlcAT-I           HLVVLTPKAQRLREG--EPGWVHPRGVEQRNKALDWLRGRGGAVGGEKDPPPPGTQGVVY  (191)
HsGlcAT-P           HLHVETPRNYKLRGDARDPRI--PRGTMQRNLALRWLR--------ETFPRNSSQPGVVY  (192)
HsGlcAT-S           HLHVPTPRRYK------RPGL--PRATEQRNAGLAWLR--------QRHQHQRAQPGVLF  (182)
                    ** * **.  .        *    **.. *** .* ***        :     .  **::
                          β4'             β5
                    E-----EEHHHHHHHHH----EEE--EE------EEEEEEE---EEEEEE----------
HsGlcAT-I           FADDDNTYSRELFEEMRWTRGVSVWPVGLVGGLRFEGPQVQD-GRVVGFHTAWEPSRPFP  (250)
HsGlcAT-P           FADDDNTYSLELFEEMRSTRRVSVWPVAFVGGLRYEAPRVNGAGKVVGWKTVFDPHRPFA  (252)
HsGlcAT-S           FADDDNTYSLELFQEMRTTRKVSVWPVGLVGGRRYERPLVEN-GKVVGWYTGWRADRPFA  (241)
                    ********* ***:*** ** ******.:*** *:* * *:. *.***: * : . ***.
                          β6   α4                  α5          β7          β7'
                    -HHH-EEEEHHHHHHH------------HHHHHHHH---HHHHHEEHHHHHH---EE--
HsGlcAT-I           VDMAGFAVALPLLLDKPNAQFDSTAPR-GHLESSLLSHLVDPKDLEPRAANCTRVLVWHT  (309)
HsGlcAT-P           IDMAGFAVNLRLILQRSQAYFKLRGVKGGYQESSLLRELVTLNDLEPKAANCTKILVWHT  (312)
HsGlcAT-S           IDMAGFAVSLQVILSNPKAVFKRRGSQPGMQESDFLKQITTVEELEPKANNCTKVLVWHT  (301)
                    :******* * ::*...:* *.  . . *  **.:*  :.  ::***.* ***.:*****
                             αΩ
                    ------HHHHHHHHHH----------
HsGlcAT-I           RTEKPKMKQEEQLQRQGRGSDPAIEV  (335)
HsGlcAT-P           RTEKPVLVNEGKKGF----TDPSVEI  (333)
HsGlcAT-S           RTEKVNLANEPKYHL----DTVKIEV  (323)
                    ****  : :* :          :*:
```

b

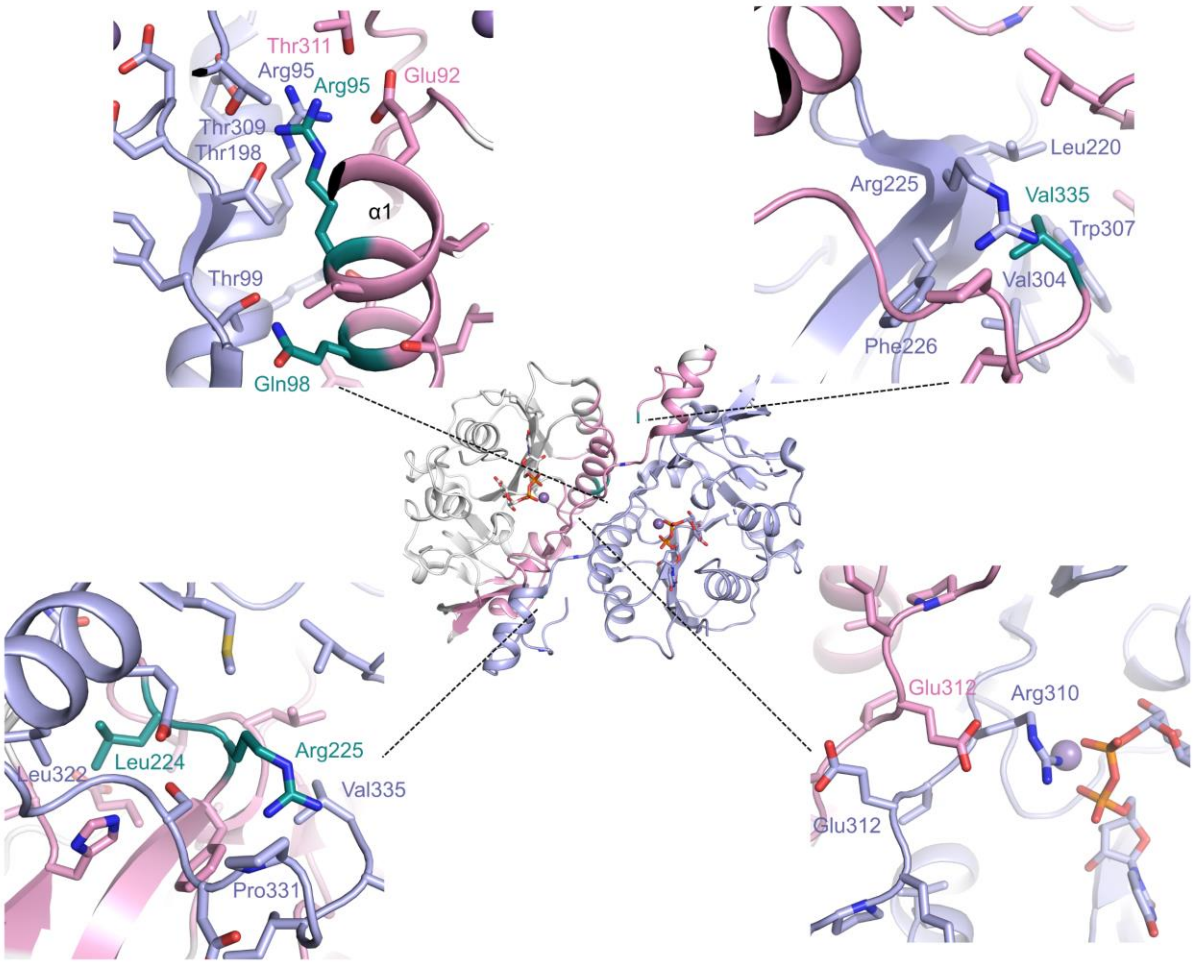


Figure 3.13 Residues involved in the homodimerisation of the catalytic domains of human GT43 β-glucuronosyltransferases. The previously solved crystal structures of GlcAT-I, GlcAT-P, and GlcAT-S were submitted to PPCheck (http://caps.ncbs.res.in/ppcheck/help.html) in order to determine the residues most likely to have a strong involvement in homodimerisation. **a** MUSCLE sequence alignment of the three human GT43 enzymes. Interface residues are highlighted in pink; 'hotspot' residues (from the top nine most energetically important interactions) are highlighted in teal. The secondary structure of GlcAT-I is annotated above the alignment; H = α-helix, E = β-sheet. **b** Structure of GlcAT-I (PDB: 1KWS); chain A is coloured as in **a**. Inter-chain interactions of the hotspot residues/salt bridge are shown in close-up panels; the opposing protomer is shown in pale blue.

Subsequently, I aligned an evolutionarily diverse subset of plant GT43 sequences to the human GT43s (which belong to the GT43-D clade) in order to see whether these interface residues could be found in the former. Plant sequences were grouped into GT43-A and GT43-B clades as in *Section 3.2.1*. Inspection of the alignment revealed that, interestingly, several of the residues involved in homodimerisation of the human GT43s were conserved in the plant sequences. For example, an analogue for Arg95 was found in every GT43-A sequence tested; Arg136 is the corresponding residue in *At*IRX9, for instance (**Figure 3.14**). In stark contrast, arginine was completely absent at this site in GT43-B sequences. Conversely, whereas amino acids resembling the Arg310–Glu312 salt bridge were well conserved in GT43-B sequences (Arg425 and Glu427 in IRX14, for instance), GT43-A2 and GT43-A3 (IRX9-related sequences) lacked such a pattern. Since the C-terminus of plant GT43 sequences is highly dissimilar to that of GT43-D sequences, however, I could not investigate the conservation of Leu224, Arg225, and Val335 analogues in the former. Nevertheless, these results support the notion that the globular domains of plant GT43s might homodimerise in a similar way to those of human GT43 enzymes. Furthermore, these results do *not* support the notion that IRX9 and IRX14 constitutively heterodimerise using this interface in a pseudosymmetric fashion. This is because such an interaction would presumably leave both Arg136 in IRX9 and Glu427 in IRX14 unpaired, giving these residues no reason to be so strongly conserved in GT43-A and GT43-B sequences, respectively.

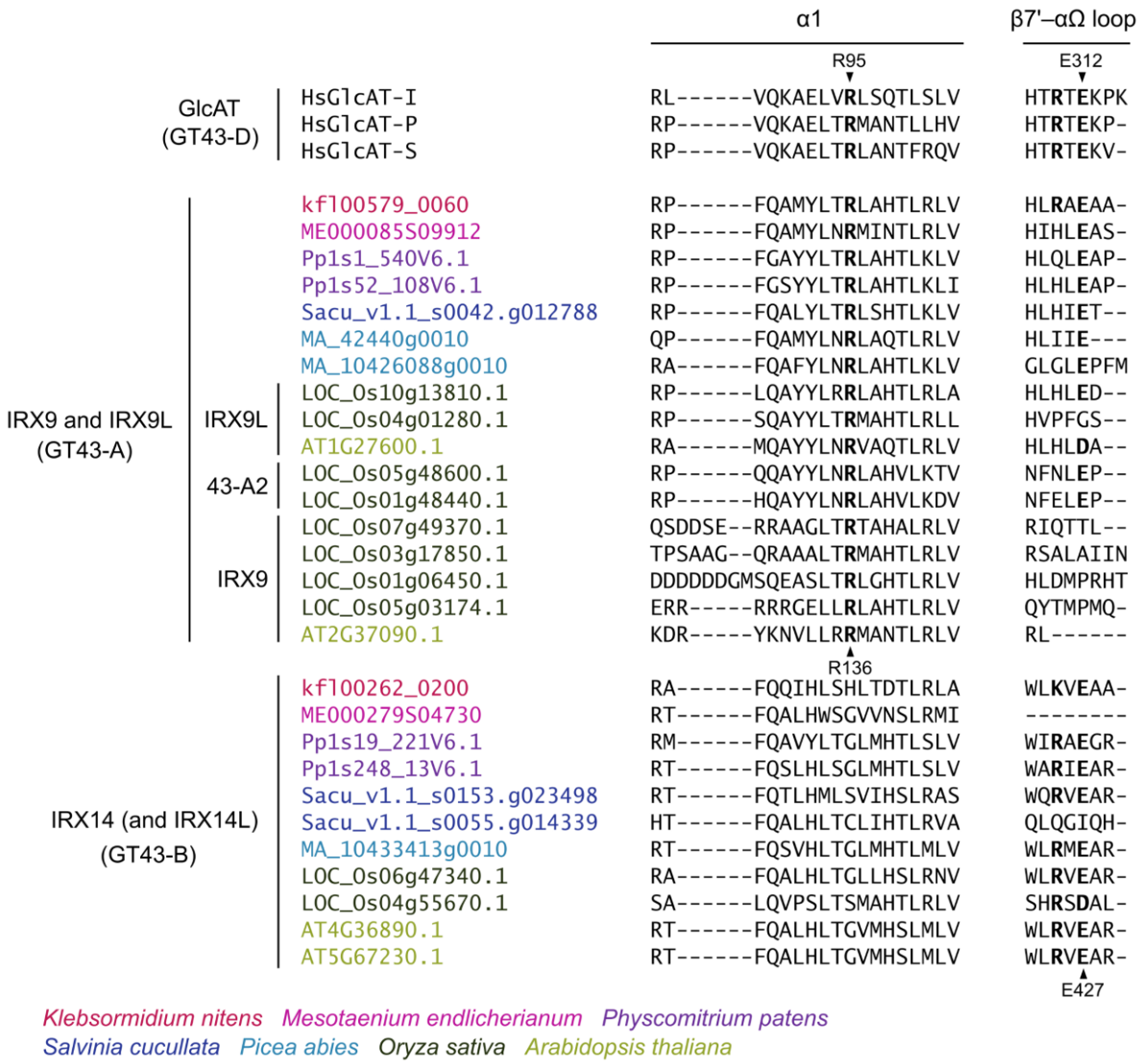


Figure 3.14 Residues involved in GT43-D homodimerisation can also be found in plant GT43 sequences. Plant GT43 sequences, sorted into GT43-A and GT43-B groups as described above, were aligned to the sequences of the three human GT43 enzymes using MUSCLE. The alignment was then truncated to the breadth of the α1 helix (containing Arg95 and Gln98) or β7'–αΩ loop (containing Arg310 and Glu312) in GlcAT-I.

3.3 Discussion

Golgi glycosyltransferases have long been known to form oligomers (Young, 2004; Hashimoto *et al*, 2010; Kellokumpu *et al*, 2016). However, such oligomerisation comprises at least two main types of interaction: symmetrical homodimerisation of individual enzymes, and the formation of hetero-complexes (typically between enzymes from the same pathway). The formation of such hetero-complexes has been proposed to permit substrate channeling and prevent cross-pathway interference within the Golgi lumen (Young, 2004; de Graffenried & Bertozzi, 2004; Oikawa *et al*, 2013; Kellokumpu *et al*, 2016). On the other hand, although GT homodimerisation has been shown to affect the localisation and/or activity of some enzymes (Young, 2004; Tu & Banfield, 2010), its function remains to be completely elucidated.

Based on current interpretations of the experimental data, some GTs appear to either participate in both types of interaction simultaneously, or switch from one to the other as they progress through the secretory pathway (Oikawa *et al*, 2013; Kellokumpu *et al*, 2016). The GT43-family enzymes IRX9 and IRX14, proposed members of the Arabidopsis XSC, are examples of such enzymes. However, a clear picture of the stoichiometry or assembly of the XSC components is currently lacking. Therefore, insight into the oligomerisation of these enzymes can provide information concerning the make-up of the XSC, as well as the wider mechanisms and functions of Golgi GT oligomerisation.

Preliminary investigations revealed the presence of highly conserved residues in the predicted TMHs of plant GT43s, as previously noted for IRX14 homologues by Jiang *et al* (2016). In this work, I examined these motifs more closely. In GT43-A sequences, which include IRX9 and IRX9L, a FxxGxxxG motif was identified. GxxxG is a well characterised transmembrane dimerisation motif (Russ & Engelman, 2000; Teese & Langosch, 2015); furthermore, the presence of phenylalanine at the −3 position relative to this motif is known to strengthen such interactions (Unterreitmeier *et al*, 2007). In GT43-B sequences, which include IRX14 and IRX14L, a CxxSxxxGxxxS motif was identified, which could also constitute an oligomerisation motif. The predicted transmembrane region of the GT43-B sequences was particularly curious, constituting an unusually long stretch of hydrophobic residues, yet with a number of conserved arginine residues found within the predicted TMH segment. Based on examples in the literature, these potentially cationic sidechains could be involved in cation-π between the two interacting helices or in sidechain 'snorkelling' that could result in helix tilting (Segrest *et al*, 1990; Johnson *et al*, 2007).

However, in itself, the presence of conserved GAS_{right} motifs such as GxxxG and SmxxxSm in a transmembrane sequence cannot be taken as an indication that the helix undergoes oligomerisation, as such amino acids can perform alternative roles in TMHs (Li *et al*, 2012; Teese & Langosch, 2015). Accordingly, I used TOXGREEN to examine the propensity of these sequences to oligomerise in *E. coli*. Maltose uptake complementation experiments indicated that for most constructs, the TMH truncation did not properly insert into the inner membrane. However, the reason for which many of the initial truncations did not insert properly is likely due to the fact that, with the downstream residues included, the hydrophobic region was too long. Nevertheless, after producing shorter truncations, I was not able to produce a construct that fully restored maltose uptake to the level of the controls. Despite this, strong GFP fluorescence was observed from cells expressing almost all of these constructs. Although full-length protein was detectable by Western blot, some apparent degradation products were visible; hence, it is possible that some of this GFP fluorescence could be attributed to the dimerisation of soluble ToxR. Nevertheless, it is also possible that the full-length proteins are able to activate reporter expression from a different location within the cell, perhaps inclusion bodies. This is consistent with the fact that, despite producing a substantial amount of GFP signal, the $IRX14_{36–55}$ construct appeared to generate far fewer degradation products than the GpA controls. Furthermore, the protein produced by the $IRX9_{21–34}$ construct was able to restore maltose uptake to a level close to that of the GpA controls; moreover, the resultant GFP expression could be abrogated by the G28I mutation, which did not appear to dramatically affect expression or degradation. Curiously, this mutation nonetheless appeared to improve the membrane integration somewhat. It is therefore possible that the oligomerisation itself has a role in protein aggregation or mislocalisation in this experimental system. Further truncations will need to be trialled in order to confirm these results; nevertheless, they suggest that the transmembrane helix of at least IRX9 (if not also IRX14) is capable of homo-oligomerisation. Therefore, it would be interesting to make further point mutants of the IRX9 TMH to test the role of other residues in dimerisation. Furthermore, IRX14 truncation trials should be continued in order to find a suitable TMH length for proper membrane insertion.

GAS_{right} motifs are found not only in homodimers, but also in trimers, tetramers, pentamers, hexamers, and heptamers, as well as heterodimers (Kim *et al*, 2005; Teese & Langosch, 2015). The TOXGREEN technique used in this work was only able to detect generic homo-oligomerisation. Hence, these results do not distinguish homodimers from higher-order structures; furthermore, heterodimerisation between the IRX9 and IRX14 TMHs was not

tested. However, further support for the homodimerisation of the IRX9 TMH lies in the fact that a homodimeric structure can be modelled with high confidence using the transmembrane dimerisation prediction tool CATM (Mueller *et al*, 2014; Anderson & Senes, 2018, pers. comm.). This is consistent with my conclusion that the globular domains of IRX9 and IRX14 also form homodimeric interactions.

To my knowledge, GAS_{right} motifs have not been documented before in Golgi glycosyltransferases. Nevertheless, precedent appears to exist in a small number of relevant ER proteins. For instance, the human ER xylosyltransferase XXYLT1 contains an AxxxAxxxA motif in its transmembrane domain that is suggested to mediate homodimerisation (Sethi *et al*, 2012). The homodimeric TMH of this enzyme has been compared to that of the human ER-localised C1GalT1 chaperone Cosmc (itself a member of GT31) (Sun *et al*, 2011; Sethi *et al*, 2012), whose predicted TMH contains a GxxxGxxxC motif (Hanes *et al*, 2017) (though this motif has not been specifically mentioned in the literature). Recently, the first functional GAS_{right} motif was identified in plants: an SxxxG motif in the ER-localised cytokinin oxidase/dehydrogenase CKX1 (Niemann *et al*, 2018). The fact that these examples are all ER-localised is interesting given that I found IRX9 to be retained in the *N. benthamiana* ER in the absence of IRX10 and IRX14 expression—as found previously for *Ao*IRX9 in the absence of *Ao*IRX10 and *Ao*IRX14, *Ao*IRX14 in the absence of *Ao*IRX9 and *Ao*IRX10, and *Ta*GT43-4 in the absence of *Ta*GT47-13 (Zeng *et al*, 2016; Jiang *et al*, 2016). However, GAS_{right} motifs need not be confined to ER-localised GTs: the (AxxxG motif-containing) TMH of Golgi-localised chondroitin sulphate glucuronosyltransferase CHPF2 currently features as the second most highly scoring homodimeric structure prediction on the CATM server homepage (http://seneslab.org/cgi-bin/CATM).

Putting aside GAS_{right} motif-driven interaction, general disulphide-linked TMH dimerisation has been frequently reported in Golgi GTs. Furthermore, removal of such disulphides through the mutation of cysteine residues, or the alteration of other interface residues, has been reported to alter localisation due to loss of disulphide-linked dimerisation (Aoki *et al*, 1992; Sousa *et al*, 2003). Interestingly, the TMH of Arabidopsis GnTI also contains a glutamine residue whose mutation affects both localisation and dimerisation (Schoberer *et al*, 2019b). Although human GT43 homodimers have not been reported to possess disulphide-linked TMHs, GlcAT-I has been shown to form intermolecular disulphide bonds in its stem region (Ouzzine *et al*, 2000). Hence, I hypothesised that the conserved cysteines found in plant GT43 TMHs might also form disulphide bridges, with a potential role in localisation. Therefore, I predicted that mutations

to the FxxGxxxG motif and conserved cysteine residue in the IRX9 TMH could have profound effects on the ability of this protein to carry out its function. Interestingly, the [G28I] mutation, which alters FxxGxxxG to FxxIxxxG, largely nullified IRX9 function. Furthermore, IRX9[G28I]-GFP appeared to be mislocalised in *N. benthamiana* leaves. However, in contrast to TMH mutants of Golgi glycosyltransferases β4GalT1 and FucT-III, which exhibit reduced Golgi retention and increased plasma membrane localisation (Aoki *et al*, 1992; Sousa *et al*, 2003), IRX9[G28I]-GFP, at least when visible, appeared to localise to large endomembrane aggregates. Because the mannosidase I marker was also re-localised to these aggregates, these results suggest that the over-expression of IRX9[G28I]-GFP somehow interferes with normal membrane trafficking. A superior experiment might have involved monitoring the localisation of IRX9[G28I] expressed at more physiological levels in Arabidopsis. I tried to express IRX9-GFP and its mutant forms under the IRX14 promoter in *irx9* Arabidopsis; however, all transformants died shortly after germination (data not shown). It is not fully clear whether this was due to a biological or a technical reason. Hence, this experiment should be re-attempted—perhaps using the IRX9 promoter in place of IRX14.

Unexpectedly, however, mutation of Cys24 to serine appeared not to affect IRX9 function, as IRX9[C24S] was able to fully complement the phenotype of *irx9*. Furthermore, I was not able to detect strong evidence for the mislocalisation of IRX9[C24S]-GFP. Hence, disulphide bond formation in the IRX9 TMH may not be necessary for function or localisation. Therefore, if a disulphide linkage is present, it might merely stabilise the more critical interaction mediated by the GAS_{right} motif. This outcome is similar to that seen for ST6GalI: although mutation of the cysteine residue in the TMH of ST6GalI results in loss of an intramolecular disulphide bridge, it does not in itself affect the localisation of this protein (Qian *et al*, 2001). Nonetheless, the potential for covalent disulphide crosslinking of the IRX9 TMH is important to investigate, as this modification would presumably maintain IRX9 as a constitutive dimer (unless the interaction can be modulated by redox conditions or *S*-acylation of these cysteines[3]): from this would come ramifications for the assembly and stoichiometry of the XSC. For this reason, I tried to detect the Myc-tagged IRX9 protein from my transgenic lines by Western blot following SDS-PAGE under reducing and non-reducing conditions. Unfortunately, however, I was not able to detect any relevant signal in this experiment; nor, strangely, was I able to detect signal when I repeated the experiment on either leaves or upper stem of plants expressing

[3] IRX9 does not appear in either Arabidopsis *S*-acylation atlas published to data, however (Hemsley *et al*, 2013; Kumar *et al*, 2020).

IRX9-Myc under the viral 35S promoter (data not shown). Clearly, further technical troubleshooting will be required to address these problems.

Nevertheless, I confirmed that IRX9[G28I] is unable to rescue the phenotype of the *irx9* mutant. The results of the monosaccharide analysis were consistent with a lack of secondary cell wall xylan synthesis in IRX9[G28I]-expressing plants, suggesting that IRX9[G28I] cannot function in xylan synthesis—though it is possible that the lack of xylan in these plants is itself due to the dwarfing phenotype. However, I was not able to establish the mechanism by which the G28I mutation disrupts the functionality of IRX9. The results from the *N. benthamiana* transient over-expression experiments suggested that IRX9[G28I] was not localised properly—hence, this mutation might cause the mislocalisation of the XSC, or even prevent its formation in the ER. However, because I could not detect any of the heterologously expressed proteins by Western blot, I was not able to ascertain the expression levels of the IRX9 mutants. Hence, it remains possible that, for some reason, IRX9[G28I] is not properly inserted into the ER membrane—or that it is degraded at an early stage.

Nevertheless, the fact that these transmembrane residues are so highly conserved across GT43-A and GT43-B sequences, from the very earliest diverging sequences in *K. nitens* to those characterised in Arabidopsis, *A. officinalis*, and wheat, indicates that they have an important function. Based on my results, as well as modelling experiments by Anderson & Senes (2018, pers. comm.), it seems likely that transmembrane homodimerisation is critical for IRX9 function. This interaction could have a particular function within the XSC. Interestingly, the co-operation of non-catalytic IRX9 with its homologue IRX14 is somewhat reminiscent of that between non-catalytic Cosmc and its homologue C1GalT1 in humans. In the latter interaction, Cosmc is thought to bind to and chaperone partially unfolded C1GalT1 in the ER (Ju & Cummings, 2014). Remarkably, as mentioned above, Cosmc also appears to possess a GAS_{right} motif in its predicted TMH; this is also true of C1GalT1 (Hanes *et al*, 2017). The close parallels between these two pairs of proteins warrants further attention—it is possible that insight into the animal proteins could inspire new hypotheses regarding the plant proteins and *vice versa*.

However, the importance of the GAS_{right} motif to IRX9 function is also consistent with a more general theme of functionally important Golgi GT homodimerisation. Although a common feature of eukaryotic Golgi glycosyltransferases (but, conspicuously, much less so of their bacterial homologues) (Hashimoto *et al*, 2010; Harrus *et al*, 2018), the function of homodimerisation (be it through the transmembrane, stem, or catalytic domain) is yet to be

fully explained. So far, the most promising and widely-conserved explanation appears to constitute a role of homodimerisation in localisation (de Graffenried & Bertozzi, 2004; Tu & Banfield, 2010). However, in the literature, TMH dimerisation, TMH sequence motifs, stem dimerisation, and stem sequence motifs seem to have been treated as distinct determinants for localisation—as though each would require detection by a different receptor or localisation factor (Colley, 1997; Tu & Banfield, 2010; Stanley, 2011). Furthermore, a plethora of alternative mechanisms for Golgi GT localisation have been proposed, including partitioning by TMH length, partitioning by TMH amino acid composition, binding of the TMH or cytoplasmic tail by COPI adaptors such as Vps74/GOLPH3, and direct binding of COPI to the cytoplasmic tail (Tu & Banfield, 2010; Welch & Munro, 2019). I speculate that many of these ideas could be consolidated in a single mechanism: one in which the localisation machinery (such as tetrameric Vps74) is somehow able to specifically recognise the *pairs* of identical TMHs or cytoplasmic tails that belong to homodimeric enzymes. The existence of such a mechanism could potentially explain: *a*) why Golgi GT homodimerisation is so common, *b*) why disruption of dimerisation affects localisation, *c*) why residues in GT cytoplasmic tails are often required for localisation, *d*) why amino acid composition of the TMH affects localisation, and *e*) why the CTS domain is usually sufficient for localisation. The two TMHs belonging to GT homodimers are likely to be situated in the same patch of membrane; hence, the 'lipophobic' effect (Li *et al*, 2012) within the membrane is likely to push these helices, as well as the adjacent cytoplasmic tails, together into close apposition. Consequently, active dimerisation of the TMH would not be strictly required by such a mechanism—only that the GT is dimerised sufficiently for this process to take effect. This model could also be extended to other Type II integral membrane proteins in the Golgi—for example, the Golgi phosphoprotein GOLPH2 requires only a disulphide-linked TMH dimer and a single positively charged residue in its cytoplasmic tail for correct localisation (Hu *et al*, 2011). Of course, this model could easily be an oversimplification—and, in reality, multiple factors are likely at play. Nevertheless, there appears to exist a wide scope for future experiments into the function and prevalence of Golgi GT TMH dimerisation, and it seems that investigation in this area will likely improve our general understanding of the mechanisms of endomembrane targeting.

Chapter 4 : Structure and activity of exostosin-like 3

4.1 Introduction

With the exception of only the most basal metazoans, heparan sulphate (HS) is ubiquitous in animal species (Yamada *et al*, 2011). This protein-linked polysaccharide has many important roles in the extracellular matrix, where it binds to a wide range of effectors according to its precise pattern of molecular decoration (Bernfield *et al*, 1999; Esko & Selleck, 2002; Sarrazin *et al*, 2011; Couchman & Pataki, 2012).

The heparan backbone is initially synthesised as an alternating polysaccharide of GlcA and GlcNAc in the Golgi apparatus by glycosyltransferases from the exostosin family: namely EXT1, EXT2, EXTL1, EXTL2, and EXTL3 in humans (Busse-Wicher *et al*, 2014). These enzymes generally have two glycosyltransferase domains: a GT47 domain, which in some exostosins harbours glucuronosyltransferase activity, and a GT64 domain, which usually possesses *N*-acetylglucosaminyltransferase activity (Kitagawa *et al*, 1999; Wei *et al*, 2000; Edvardsson *et al*, 2011; Geshi *et al*, 2011). Whereas the EXT1:EXT2 hetero-complex is mainly responsible for backbone elongation, EXTL3 is likely the principal enzyme involved in backbone initiation, which constitutes the addition of the first GlcNAc residue to the tetrasaccharide linker (GlcNAcT-I activity; see **Figure 4.1**) (Kim *et al*, 2001; Yamada, 2020). Although EXTL3 also appears to add GlcNAc residues during the extension phase (GlcNAcT-II), this enzyme is not thought to possess any glucuronosyltransferase activity (GlcAT-II) (Kim *et al*, 2001).

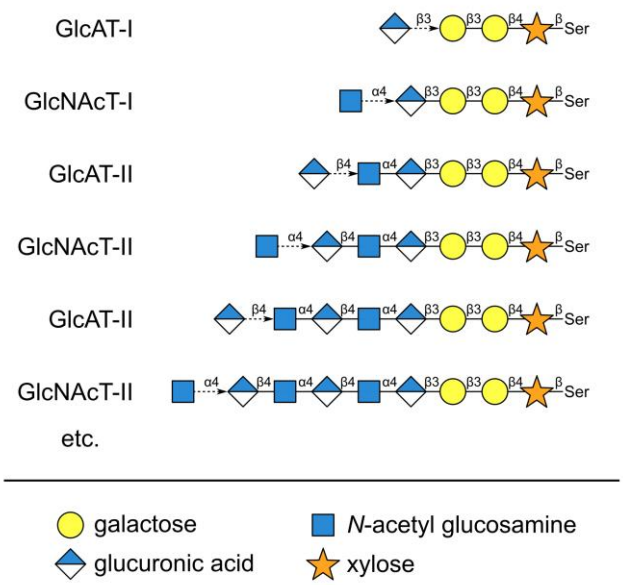


Figure 4.1 GlcAT and GlcNAcT activities in heparan sulphate backbone synthesis.

EXTL3 constitutes the largest of the exostosins, with a GT47 domain, a GT64 domain, and a predicted coiled-coil domain in the stem region (Zak *et al*, 2002; Awad *et al*, 2018). EXTL3 represents an interesting target for structural characterisation for at least two reasons: because

it contains more than one glycosyltransferase domain—and could therefore give valuable insight into how glycosyltransferase activities are physically arranged in the Golgi—and because it contains a GT47 domain, which could give equally valuable insight into the structures of related enzymes in plants (which include IRX10 and a large variety of other cell wall-synthesising activities). This latter aspect could provide structural models of plant GT47 enzymes, potentially facilitating the rationalisation and prediction of substrate specificities in this family. Furthermore, its predicted symmetry makes it an easier target for study than the potentially pseudosymmetric EXT1:EXT2 complex.

In this chapter, I describe an updated method for assaying heparan backbone synthase activities and report high-resolution cryo-EM structures for both apo-EXTL3 and UDP-bound EXTL3. I show that the GT47 domain of EXTL3 has likely lost activity through extension of the Cα4 helix and the formation of a salt bridge between Glu453 and Arg421. I also describe a hitherto unreported clade of GT47-family enzymes from lower plants that are closely related to animal exostosins. As well as providing insight into the functioning of bi-domain glycosyltransferase and nucleotide sugar binding in GT-B enzymes, these results lay down a path to further experiments on plant GT47 enzymes.

Contribution note: Many aspects of this project were collaborative in nature. Expression of EXTL3ΔN was performed by Katrin Mani (Lund University). Purification of EXTL3ΔN and obtainment of the apo-EXTL3 cryo-EM map were achieved in close collaboration with Steven Hardwick (University of Cambridge) and Tom Dendooven (formerly University of Cambridge, now moved to MRC Laboratory of Molecular Biology, Cambridge). A second batch of EXTL3ΔN was purified by the Lund Protein Purification Platform. Collection of MALDI-TOF mass spectra was performed by Theodora Tryfona (University of Cambridge). Further details are provided below in the main text.

4.2 Results

4.2.1 Glycosyl hydrolases Pa*GH89 and* Thar*GH79a/b exhibit* exo-*acting α-N-acetylglucosaminidase and β-glucuronidase activities against K5 heparosan, respectively*

Awad et al. (2018) recently reported a method for expressing EXTL3 in a soluble, secreted form ('EXTL3ΔN') in which the first 51 amino acids of the regular protein are replaced with a BM40 secretion peptide, a 6×His tag, and a TEV cleavage site (**Figure 4.2**). This form of EXTL3, which still contains the predicted coiled-coil domain, the GT47 domain, and the GT64 domain, represented an ideal subject for investigation due to its solubility and secretion into

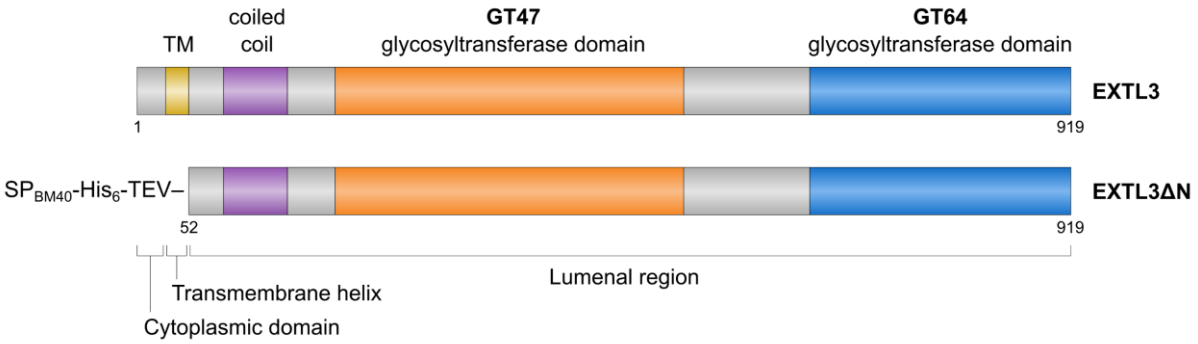


Figure 4.2 Schematic showing domain structure of EXTL3 and EXTL3ΔN.

the cell growth medium. However, before embarking on structural work, I wanted to assay the activity of EXTL3ΔN in order to demonstrate that this form of the enzyme is active.

However, appropriate carbohydrate acceptors are not readily available for the characterised activities of EXTL3. In particular, it was not possible to obtain a suitable analogue for the tetrasaccharide linkage region—to which EXTL3 exhibits its primary activity (GlcNAcT-I); hence, I did not test for this activity. Nevertheless, GlcNAcT-II and GlcAT-II activity assays have previously made use of the capsular polysaccharide of *E. coli* K5, which, with its structure of alternating β1,4-linked GlcA and α1,4-linked GlcNAc residues, constitutes an analogue of the nascent, unmodified heparan backbone. Typically, this polysaccharide, termed 'K5 polysaccharide', 'heparosan', or 'K5 heparosan', is partially deacetylated using hydrazine before undergoing deaminitive cleavage specifically between GlcN and GlcA residues—thus creating products with GlcA at the non-reducing terminus. Glycosyl hydrolase (GH) enzymes, on the other hand, do not seem to have been previously used to prepare or analyse heparosan acceptors. Because such enzymes would eliminate the need for these potentially hazardous and less specific chemical reagents (as well as constituting a novel tool for the analysis of GlcNAcT and GlcAT products), I first wanted to develop an enzymatic toolkit for the preparation and manipulation of heparosan oligosaccharides.

To that end, I obtained a commercially available K5 heparosan lyase product, with degree of polymerisation (DP; number of monosaccharide units) equal to 10. Owing to the lyase activity used to produce this oligosaccharide, it possessed the structure Δ4,5HexA-GlcNAc-[GlcA-GlcNAc]$_4$, where Δ4,5HexA is Δ4,5-unsaturated hexuronic acid. To identify enzymes that would be able to break the various glycosidic linkages in this structure, I incubated the oligosaccharide sequentially with several GHs, removing each enzyme by filtration between

each step; the products were subsequently derivatised with an anionic 8-aminonaphthalene-1,3,6-trisulphonic acid (ANTS) fluorophore and separated using the electrophoretic technique 'polysaccharide analysis by carbohydrate electrophoresis' (PACE). In the undigested sample, the main band was assumed to represent the DP10 Δ4,5HexA-GlcNAc-[GlcA-GlcNAc]$_4$ oligosaccharide (**Figure 4.3**). Several additional bands were visible, however: the band above the main band was determined to be a labelling artefact (caused by incomplete reduction of the labelling intermediate) and/or DP9 GlcNAc-[GlcA-GlcNAc]$_4$, while the next most intense band was deemed to constitute DP12 Δ4,5HexA-GlcNAc-[GlcA-GlcNAc]$_5$. For GH digestions, I began by assaying the activity of BT4658^{GH88} from *Bacteroides thetaiotamicron* VPI-5482: a GH88-family enzyme that has been reported to remove Δ4,5HexA from HS lyase products (Cartmell *et al*, 2017). After treatment with BT4658^{GH88}, the band corresponding to the original DP10 oligosaccharide shifted upwards in the PACE gel, consistent with the loss of

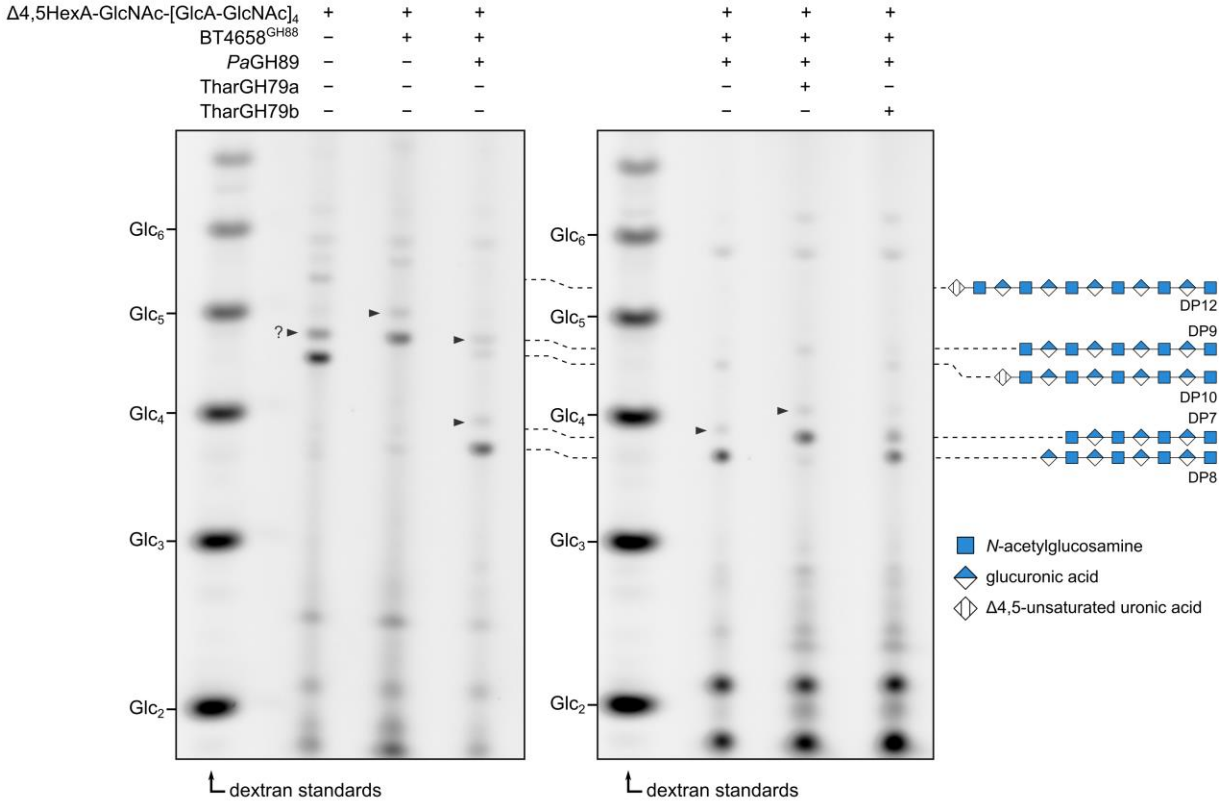


Figure 4.3 BT4658^{GH88}, *Pa*GH89, *Thar*GH79a, and *Thar*GH79b exhibit activity on heparosan lyase products. DP10 K5 heparosan lyase product was incubated with the indicated enzymes in a sequential manner. Enzymes were removed using a centrifugal filtration device with a 3 kDa cut-off. Dark grey arrows indicate labelling artefacts due to incomplete reduction of the Schiff base derivatisation intermediate (verified by emission wavelength). The standards are products from acid hydrolysis of dextran.

the negatively charged Δ4,5HexA residue (this was later confirmed with mass spectrometry—see *Section 4.2.2* below). An additional band was also present at a slightly higher position in this lane; however, this was confirmed to represent another labelling artefact (by virtue of its altered emission wavelength under illumination with UV). Next, I treated the resultant oligosaccharide (presumed to possess a DP9 GlcNAc-[GlcA-GlcNAc]$_4$ structure) with a novel fungal GH89 enzyme, referred to here as *Pa*GH89 (chosen because the GH89 family also contains the human α-*N*-acetylglucosaminidase NAGLU). Treatment with *Pa*GH89 resulted in a significant downwards shift, consistent with the loss of the uncharged non-reducing-terminal GlcNAc residue—hence, the product was deduced to exhibit the DP8 structure [GlcA-GlcNAc]$_4$. Finally, I tested two novel GH79 enzymes from *Trichoderma harzianum*, referred to here as *Thar*GH79a and *Thar*GH79b (GH79 contains both *exo*-acting β-glucuronidases as well as *endo*-acting heparanases and hyaluronidases). Similarly to GH88 treatment, *Thar*GH79a treatment resulted in a small upwards shift of the DP8 band, consistent with the loss of negatively charged GlcA. This product was deduced to possess the DP7 structure GlcNAc-[GlcA-GlcNAc]$_3$. Treatment with *Thar*GH79b resulted only in a partial shift, indicating that this enzyme was less active under the conditions used. These results not only confirm that BT4658^{GH88} is active on K5 heparosan lyase products, but also suggest that *Pa*GH89 possesses α1,4-*N*-acetylglucosaminidase activity on heparosan, and suggest that *Thar*GH79a, and to some extent *Thar*GH79b, exhibit heparosan-compatible β1,4-glucuronidase activity.

4.2.2 Preparations of EXTL3ΔN exhibit not only GlcNAcT-II activity but also an appreciable amount of GlcAT-II activity

Having established a system to characterise heparosan structures, I proceeded to test the activity of EXTL3ΔN. To this end, we purified EXTL3ΔN from culture supernatant by nickel-NTA and size exclusion chromatography. Protein was expressed in human 293-EBNA cells by Prof Katrin Mani. Protein was then purified from the cell supernatant. I purified one biological replicate in Cambridge with assistance from Dr Steven Hardwick (‘batch 1’), while a second was purified by the Lund Protein Purification Platform (‘batch 2’). To assay the purity of these preparations, I submitted them to the Cambridge Centre for Proteomics for analysis. Dr Yagnesh Umrania analysed the data by label-free quantification in order to estimate the abundances of individual proteins (derived from the peak areas of assigned peptides). The results revealed that these preparations did not contain high levels of contaminants;

furthermore, the data indicated that the levels of EXT1 and EXT2 were 1,000–10,000-fold less than that of EXTL3 (**Table 4.1** and **Table 4.2**).

Subsequently, I incubated a DP8 $[GlcA\text{-}GlcNAc]_4$ acceptor, produced by treatment of the DP10 K5 lyase product with $BT4658^{GH88}$ and *Pa*GH89, with a combination of the purified EXTL3ΔN and/or UDP-GlcNAc for 1 or 18 hours. Furthermore, because GlcNAc-T activity is thought to originate from the GT64 domains of exostosins, which adopt a metal-dependent GT-A fold, I also tested the metal-dependence of the activity by replacing $MnCl_2$ and $MgCl_2$ in the buffer with EDTA. The products were derivatised with ANTS (with a higher concentration of reducing agent in order to reduce the abundance of labelling artefacts) and separated by PACE. Judging by the migration of the resultant bands, incubation of the DP8 acceptor with either EXTL3ΔN or UDP-GlcNAc alone did not result in any new product, whereas overnight incubation with both enzyme and donor substrate resulted in the complete conversion of the DP8 oligosaccharide to DP9 GlcNAc-$[GlcA\text{-}GlcNAc]_4$ (**Figure 4.4a**). When the reaction was terminated after only one hour, a very small amount of product was also visible (though this was barely above the background level of DP9 oligosaccharide). In contrast, EXTL3ΔN did not exhibit any GlcNAcT-II activity in the presence of EDTA. These results confirm that, as expected, our preparations of EXTL3ΔN exhibit metal-dependent GlcNAcT-II activity, and that therefore, EXTL3ΔN is likely to be correctly folded in these preparations.

Table 4.1 Proteomic analysis of EXTL3ΔN batch 1. Semi-quantitative abundance, determined by label-free quantification (peptide threshold = 90%), is listed for the top ten most abundant proteins, as well as for EXT1 and EXT2.

Protein name	Accession	Abundance (a.u.)
EXTL3	O34909	1.0×10^{11}
Trypsin (*Sus scrofa*)	cRAP112	7.9×10^{9}
Calsyntenin-1	O94985	1.1×10^{9}
Fibulin-1	P23142	4.7×10^{8}
LTBP4 isoform 2	Q8N2S1-2	2.9×10^{8}
Nidogen-1	P14543	1.6×10^{8}
GFP (*Æquorea victoria*)	cRAP032	1.5×10^{8}
RAN	B5MDF5	1.5×10^{8}
AHSG (*Bos taurus*)	cRAPR1	1.2×10^{8}
PXDN	Q92626	1.1×10^{8}
…		
EXT2	Q93063	1.2×10^{7}
…		
EXT1	Q16394	1.6×10^{6}

Table 4.2 Proteomic analysis of EXTL3ΔN batch 2. Semi-quantitative abundance, determined by label-free quantification (peptide threshold = 90%), is listed for the top ten most abundant proteins, as well as for EXT1 and EXT2.

Protein name	Accession	Abundance (a.u.)
EXTL3	O34909	1.0×10^{12}
Trypsin (*Sus scrofa*)	cRAP112	3.2×10^{10}
GLUD1	P00367	2.1×10^{10}
Actin, cytoplasmic 1	P60709	2.6×10^{9}
Clusterin	P10909	2.5×10^{9}
HSPA1B	A0A0G2JIW1	2.5×10^{9}
HSP90AA1	P07900	2.0×10^{9}
HTRA1	Q92743	1.6×10^{9}
LDHB	P07195	1.6×10^{9}
LDHA	P00338	1.5×10^{9}
…		
EXT2	Q93063	1.2×10^{8}
…		
EXT1	Q16394	3.9×10^{7}

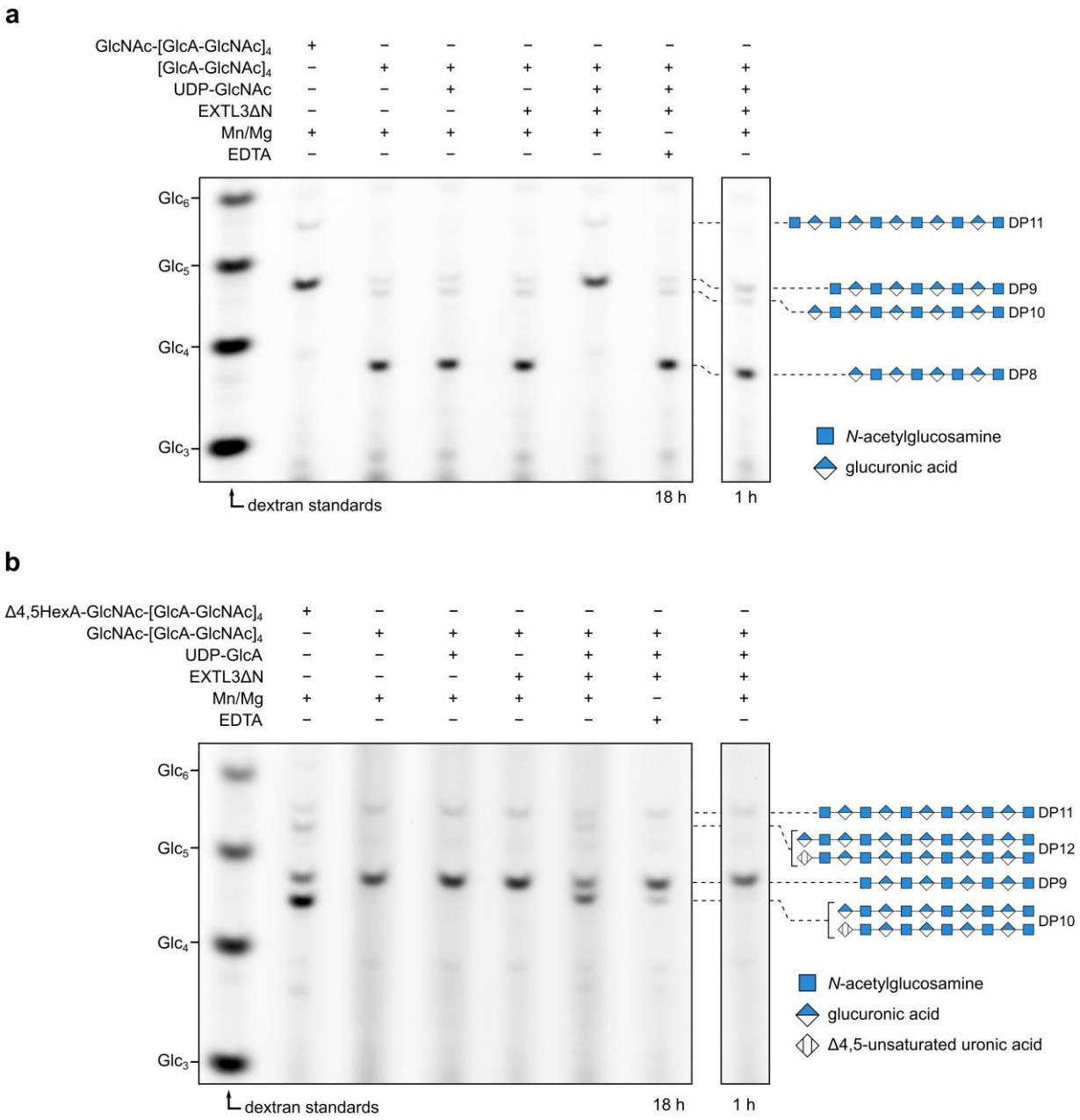


Figure 4.4 Preparations of EXTL3ΔN appear to exhibit both GlcNAcT-II and GlcAT-II activity. Acceptor oligosaccharides were prepared by GH digestion of a DP10 K5 heparosan lyase product. Acceptor was incubated for 1 or 18 h with EXTL3ΔN and/or UDP-sugar (1.5 mM) at 37 °C before derivatisation of the products with ANTS and analysis by PACE. **a** GlcNAcT-II activity; representative of EXTL3ΔN batch 1 and batch 2. **b** GlcAT-II activity; representative of EXTL3ΔN batch 2 (see **Figure 4.5** for experiment with batch 1).

EXTL3 is not thought to possess GlcAT-II activity (Yamada, 2020). Nevertheless, because preliminary data from my collaborators suggested that EXTL3 might have a low level of GlcAT-II activity (Mani & Logan, 2018, pers. comm.), I assayed for this activity in a similar manner to that for GlcNAcT-II activity: I incubated a DP9 GlcNAc-[GlcA-GlcNAc]$_4$ acceptor, produced by treatment of the DP10 K5 lyase product with BT4658^{GH88}, with a combination of the purified EXTL3ΔN and/or UDP-GlcA in an overnight reaction. Once again, I tested the metal-dependence of the activity using EDTA. The products were derivatised with ANTS and separated by PACE. Surprisingly, incubation with both EXTL3ΔN and UDP-GlcA produced a new band at the position expected of DP10 [GlcA-GlcNAc]$_5$ (**Figure 4.4b**). However, unlike the GlcNAcT-II reaction, this reaction did not appear to reach completion after 18 h. Interestingly, this activity was only partly inhibited by EDTA, suggesting that it was unlikely to originate from the metal ion-dependent GT64 domain.

Because this activity would be somewhat controversial for EXTL3, I sought to characterise the apparent DP10 product further. Firstly, to confirm the donor sugar specificity of this reaction, I incubated the DP9 acceptor with EXTL3ΔN and UDP-GlcNAc. No activity was seen in this case, indicating that GlcNAc cannot be transferred directly to the DP9 acceptor (**Figure 4.5**). Secondly, I treated the DP10 product with two different β-glucuronidases: a GH2 β-glucuronidase from bovine liver, and *Thar*GH79a. In both cases, the DP10 product was sensitive to enzyme digestion although the GH2 digestion did not appear to progress to completion under the conditions of this experiment. Nevertheless, this result constitutes strong evidence that the new moiety is a β-linked glucuronosyl residue. Finally, I provided EXTL3ΔN with both UDP-GlcA and UDP-GlcNAc (in equimolar concentrations). Remarkably, a ladder of products was seen, which likely constituted higher-order heparosan oligosaccharides produced by alternating GlcAT-II and GlcNAcT-II activities. Together, these results strongly support the hypothesis that our EXTL3ΔN preparations exhibit GlcAT-II activity.

To gather further evidence for the identity of these oligosaccharides, the relevant products were characterised by matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry (MS). Prior to analysis, I desalted the oligosaccharides and derivatised them with 2-aminobenzamide (2-AB). Mass spectrometry was then performed by Dr Theodora Tryfona. For the undigested DP10 heparan lyase product, an ion with m/z ratio of 2014.9, which corresponds to the mass of deprotonated ΔHexAGlcNac$_5$GlcA$_4$-2-AB, was detected following the MALDI-TOF analysis (**Figure 4.6**). A smaller ion of 1856.7 m/z (GlcNac$_5$GlcA$_4$-2-AB) was also detected, and presumably indicates partial hydrolysis of the Δ4,5HexA residue. After

treatment with BT4658^{GH88}, the peak at 2014.9 m/z disappeared, whereas the 1856.7 m/z peak became much stronger, confirming that this enzyme completed the removal of the non-reducing-terminal Δ4,5HexA residue. When this DP9 oligosaccharide was incubated with EXTL3ΔN and UDP-GlcA, this peak was shifted to 2032.7 m/z, indicating the gain of a (saturated) hexuronic acid, presumably glucuronic acid. Furthermore, after incubation with EXTL3ΔN, UDP-GlcA, and UDP-GlcNAc, the main peaks observed were at 1856.6 m/z ($GlcNac_5GlcA_4$-2-AB), 2235.7 m/z ($GlcNac_6GlcA_5$-2-AB), 2615.8 m/z ($GlcNac_7GlcA_6$-2-AB), 2994.9 m/z ($GlcNac_8GlcA_7$-2-AB), 3374.0 m/z ($GlcNac_9GlcA_8$-2-AB), and 3753.1 m/z

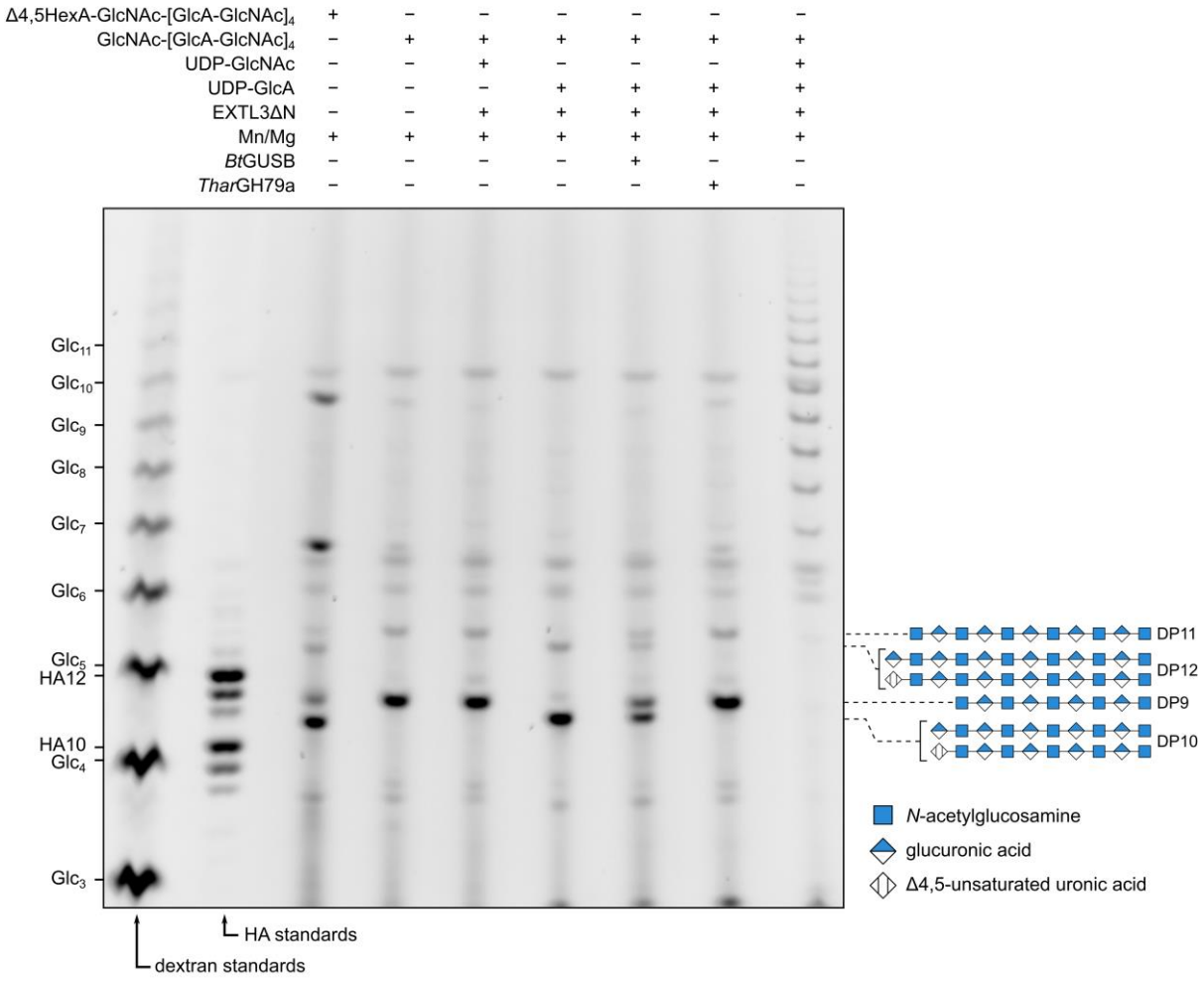


Figure 4.5 Confirmation of GlcAT-II activity. DP9 acceptor oligosaccharide was prepared by digesting a DP10 K5 heparosan lyase product with BT4658^{GH88}. Acceptor was incubated for 18 h with EXTL3ΔN (batch 1) and/or UDP-sugar (1.5 mM) at 37 °C. After termination of the reactions, two of the samples were subjected to an overnight incubation with either GH2 β-glucuronidase from bovine liver (*Bt*GUSB) or *Thar*GH79a. All samples were then derivatised with ANTS and separated by PACE.

($GlcNac_{10}GlcA_9$-2-AB). A second series of peaks, including 2056.9 m/z ($GlcNac_5GlcA_5$-2-AB), 2412.7 m/z ($GlcNac_6GlcA_6$-2-AB), 2791.8 m/z ($GlcNac_7GlcA_7$-2-AB), 3169.9 m/z ($GlcNac_8GlcA_8$-2-AB), and 3551.0 ($GlcNac_9GlcA_9$-2-AB), was also visible, but the intensities of these peaks were much weaker in comparison. These results demonstrate that, in the presence of equimolar UDP-GlcA and UDP-GlcNAc, the products produced by EXTL3ΔN are predominantly of odd degrees of polymerisation. Assuming that these oligosaccharides have an alternating GlcNAc and GlcA structure, this implies that GlcNAc is transferred to the non-reducing terminus at a higher rate than GlcA is.

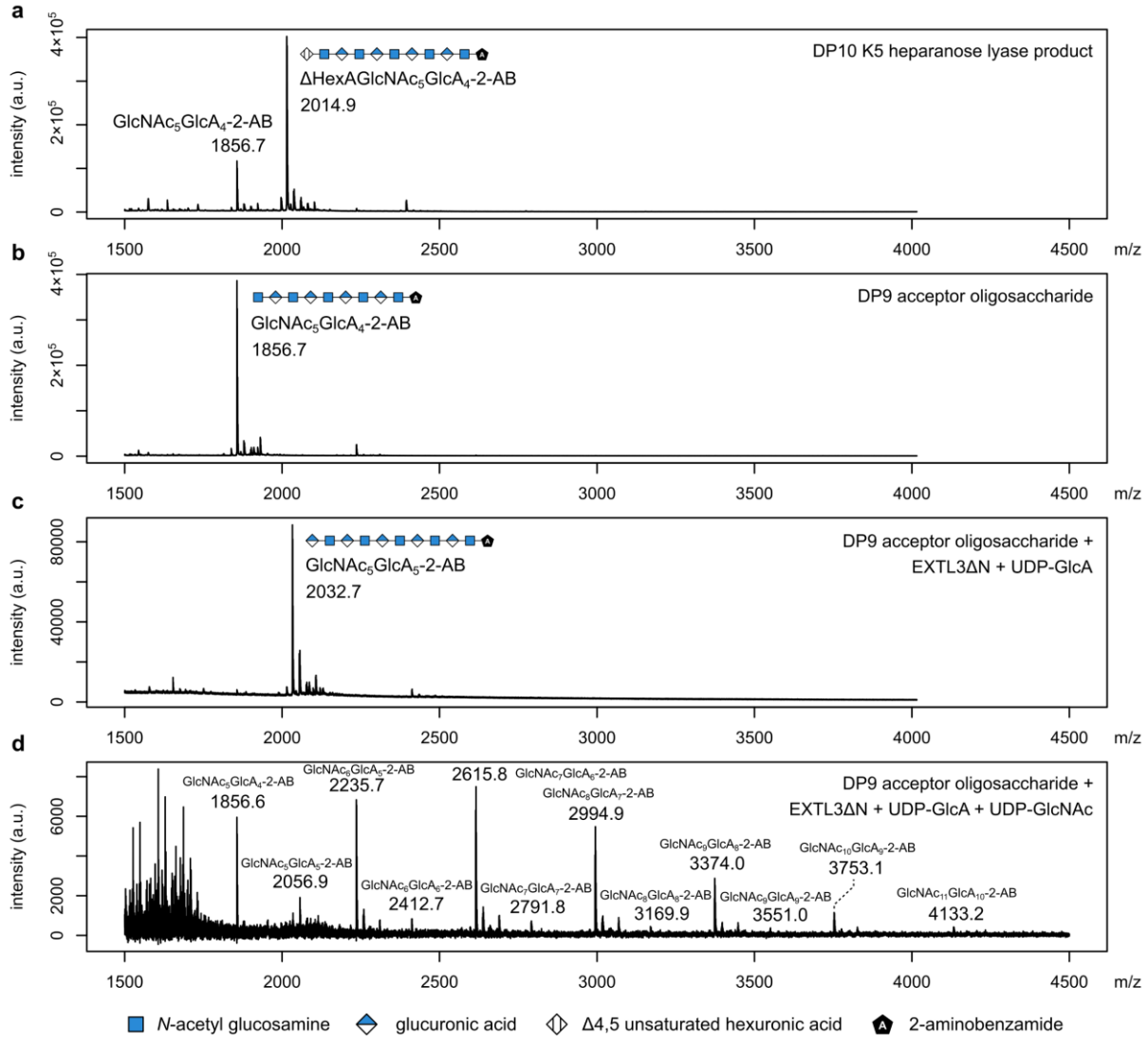


Figure 4.6 Verification of product assignments using MALDI-TOF MS. DP9 acceptor oligosaccharide were prepared by digesting a DP10 K5 heparosan lyase product with $BT4658^{GH88}$. Acceptor was incubated for 18 h with EXTL3ΔN (batch 1) and/or UDP-sugar (1.5 mM) at 37 °C. All samples were derivatised with 2-aminobenzamide prior to analysis by MALDI-TOF mass spectrometry.

4.2.3 Preliminary kinetics data suggest that the K_M for UDP-GlcA is lower than that for UDP-GlcNAc

Because a GlcAT-II activity has not been previously described for EXTL3, I wanted to compare the kinetics of the GlcNAcT-II and GlcAT-II activities. Hence, I repeated the two reactions under various UDP-sugar concentrations and reaction times. After the completion of each reaction, the concentration of liberated UDP was quantified using a commercial luciferase-based assay (in which UDP is converted to ATP). Unfortunately, due to time and material constraints, I was unable to collect a full dataset. Hence, it would not have been appropriate to calculate numerical K_M or V_{max} values for these data. Nevertheless, judging based on the shapes of the curves from the preliminary Michaelis-Menten plots (**Figure 4.7**), it appeared that, surprisingly, the K_M value for UDP-GlcA may in fact be lower than that for

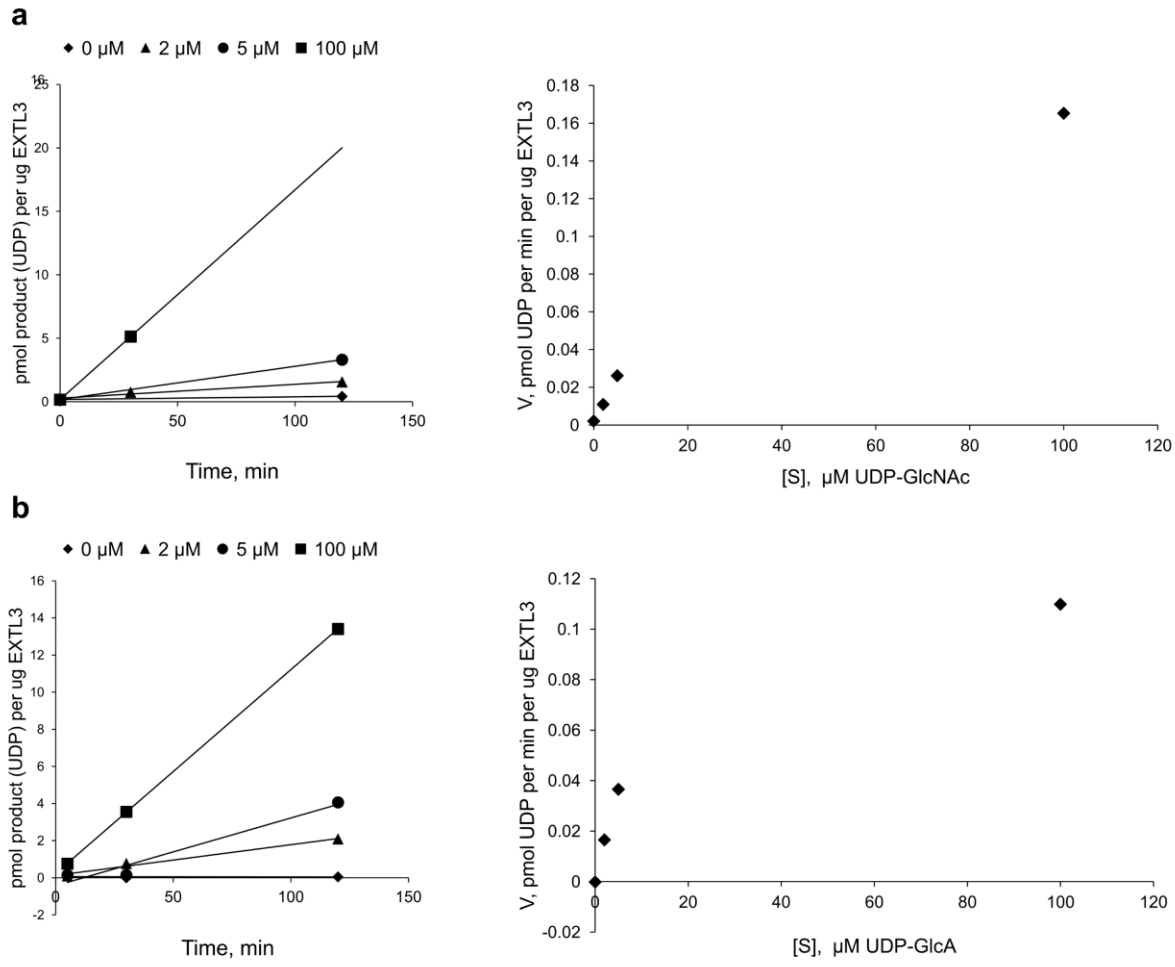


Figure 4.7 K5 acceptor was incubated with EXTL3ΔN (batch 1) and UDP-sugar at the specified concentration at 37 °C for the specified times. Released UDP was quantified using UDP-Glo luciferase-based assay. **a** GlcNAcT-II activity (DP8 acceptor and UDP-GlcNAc). **b** GlcAT-II activity (DP9 acceptor and UDP-GlcA).

UDP-GlcNAc. In addition, at 100 µM UDP-sugar concentration, the levels of the two activities were within the same order of magnitude. Therefore, these results also suggest that the GlcAT-II activity observed above (at 1.5 mM UDP-GlcA) is not merely an artefact of high UDP-GlcA concentration.

4.2.4 Single-particle cryo-EM reveals the overall structure of the EXTL3 catalytic domain at high resolution

EXTL3 is the largest of the five human exostosins. In addition to a transmembrane domain and cytoplasmic tail, EXTL3 is predicted to possess a coiled-coil domain in the stem region, a GT47 domain, and a C-terminal GT64 domain. Furthermore, EXTL3 is predicted to form a homodimer; hence, a biological assembly ought to contain two of each GT domain, and, in total, would most likely possess a molecular weight of 200 kDa.

However, within this dimer, the locations of the active sites are yet to be confirmed. Similarly, the structural arrangement of these domains, whose close association has been proposed to allow substrate channeling (Young, 2004), is unknown. Furthermore, although the EXTL3 GT47 domain has been tentatively predicted to exhibit a GT-B-type conformation (Awad *et al*, 2018), the fold adopted by enzymes in this family is yet to be confirmed (hindering the design of experiments on related plant enzymes). Hence, to gain insight into the architecture of this enzyme, we solved the structure of EXTL3ΔN (batch 1) by single-particle cryogenic electron microscopy (cryo-EM). Dr Steven Hardwick prepared cryo-EM grids from the purified protein, Dr Dima Chirgadze clipped the grids and operated the microscope, and I processed the data with assistance from Mr Tom Dendooven. I was able to produce a density map at 3.0 Å resolution; subsequently, further work by Tom Dendooven improved the resolution to 2.4 Å (**Figure 4.8**). I constructed a *de novo* co-ordinate model from this density model using Buccaneer, Coot, and Phenix Real-Space Refine. From this, it was clear that the density belonged to the globular catalytic domain of EXTL3 (**Figure 4.9**, **Figure 4.10**). The density did not appear to contain the coiled-coil domain, which was presumably masked due to its flexibility.

As expected, the catalytic domain dimer is symmetrical, and comprises four distinct glycosyltransferase-like domains. Connecting the N-terminal GT47 domain to the C-terminal GT64 domains is a substantial linker region. Containing 124 residues, this linker not only bridges the gap between the GT47 and GT64 domains, but also participates in homodimeric interactions at the core of the protein. Furthermore, the linker region forms an extended loop

over one side of the GT64 domain (**Figure 4.9b**), the purpose of which is not immediately clear.

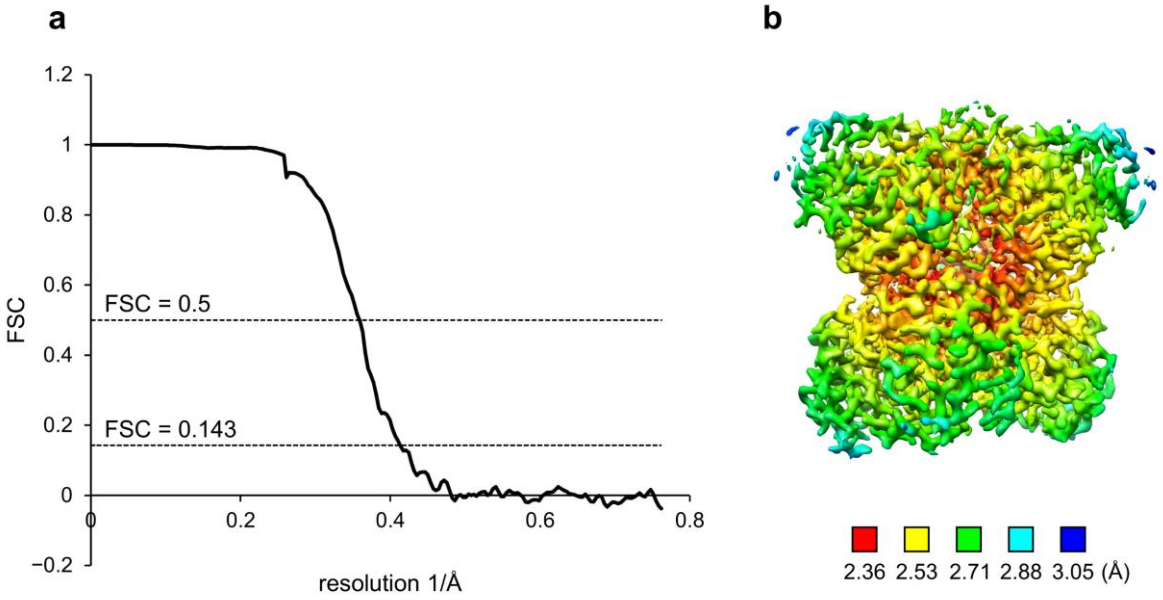


Figure 4.8 Resolution estimates for the apo-EXTL3 density map. a FSC curve. **b** Local resolution map, calculated using RELION.

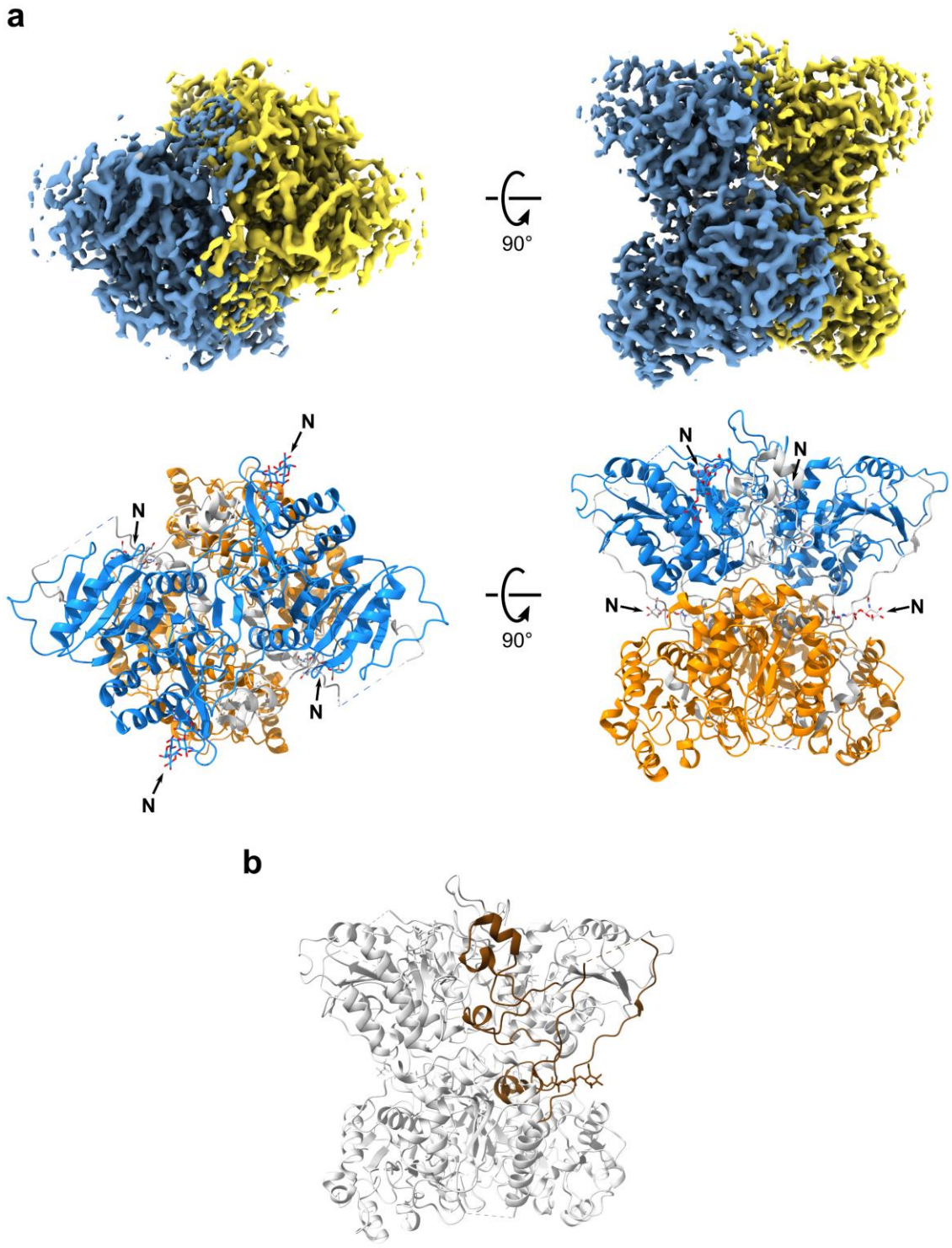


Figure 4.9 Overall structure of the apo-EXTL3 catalytic domain dimer. a Map and model. Top: cryo-EM density map at contour level 0.057 (top and side views). The two constituent monomers are shown in blue and yellow, respectively. Bottom: corresponding atomic model. GT47 domains are shown in orange; GT64 domains are shown in blue. *N*-glycans attached at Asn592 and Asn790 are shown in stick representation and labelled as **N**. **b** The linker region is shown in brown.

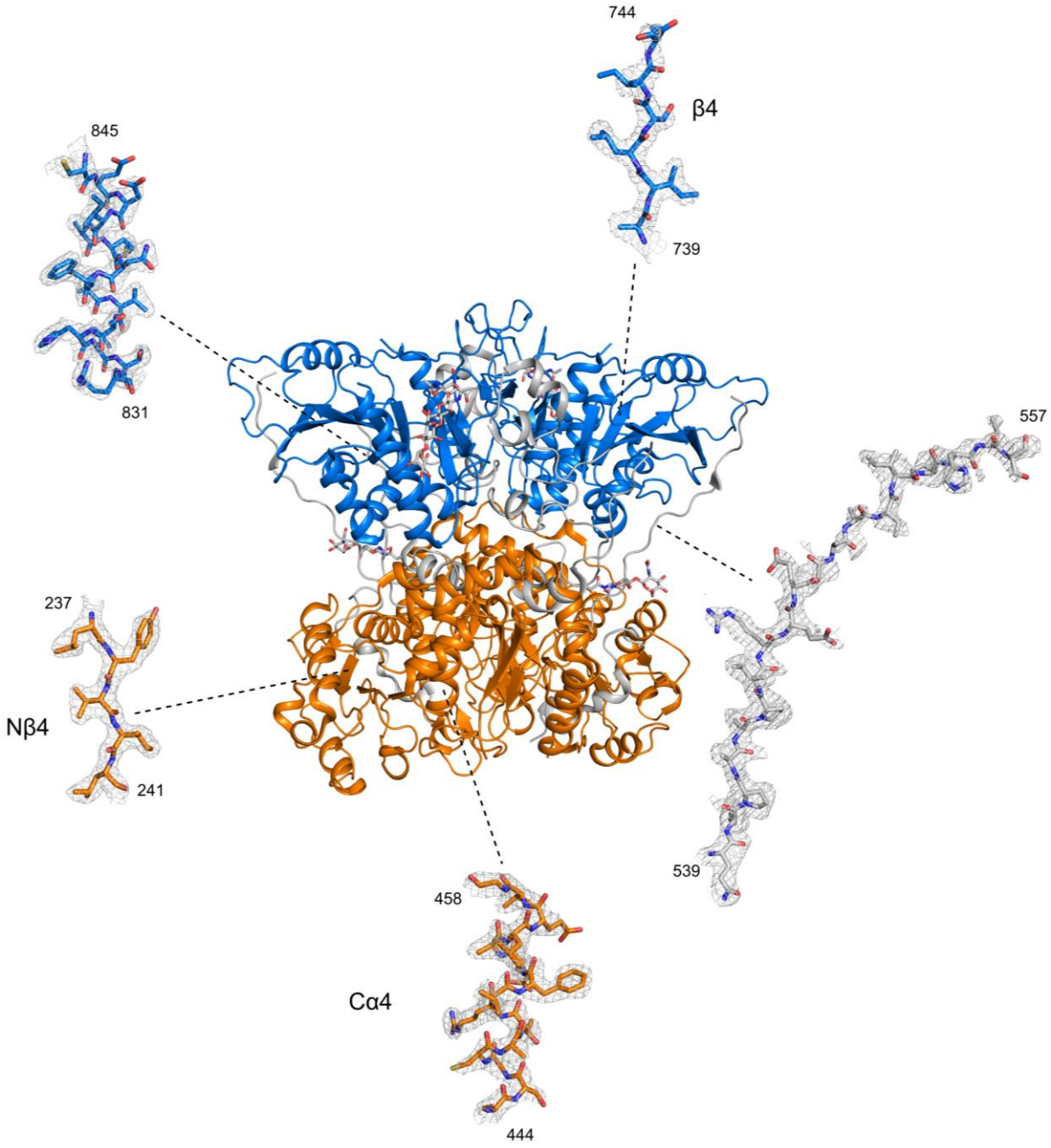


Figure 4.10 Local cryo-EM density at key areas of the apo-EXTL3 map.

4.2.5 The GT64 domain structure of EXTL3 is highly similar to that of mouse EXTL2, but the two differ in their C-termini

Previously, only one other exostosin has been structurally characterised: EXTL2 from mouse (Pedersen *et al*, 2003). As in human EXTL2, mouse EXTL2 (*Mm*EXTL2) does not possess a GT47 domain; hence, the published structure consists only of a GT64 domain homodimer. I compared the structure of an individual GT64 domain from EXTL3 to that of mouse *Mm*EXTL2 by aligning them in PyMOL and calculating an RMSD between the aligned C_α atoms. The results revealed that the two are highly similar (**Figure 4.11a**), with a calculated RMSD of 0.94 Å (between 161 aligned atoms). However, one major discrepancy is constituted by the difference in structure between the two C-termini: the C-terminal loop of EXTL3 is somewhat longer than that of *Mm*EXTL2, and, since its very terminus is rooted in the protein fold, appears also to be more structured. EXTL3 has been proposed to somehow recognise a single hydrophobic residue and multiple anionic residues in the core protein of its substrate proteoglycans (Esko & Zhang, 1996). Interestingly, I found that the C-terminus of EXTL3 has many cationic side chains, and additionally exhibits a phenylalanine residue (Phe918) whose sidechain is exposed to the solvent (**Figure 4.11b**). Therefore, I hypothesised the C-terminus could have a role in recognition of the core protein.

To explore this idea further, I used the phmmer server (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer) to identify over 300 protein sequences with homology to EXTL3 from diverse metazoan species. These sequences were aligned; subsequently, I used WebLogo 3 (Crooks *et al*, 2004) to create a sequence logo for the sites corresponding to the C-terminal loop of human EXTL3. Interestingly, this region, including the cationic and hydrophobic residues of interest, appeared to be highly conserved (**Figure 4.11c**). Hence, these results are consistent with an important role of the C-terminus in EXTL3 function.

Since both GT64 domains form symmetrical homodimers, I also compared the way in which this is accomplished in the two proteins. To achieve this, I aligned both homodimeric structures, using one chain only as the reference. The alignment showed that both GT64 domains homodimerise in a similar fashion, albeit with a small difference in the relative angle between the two subunits (**Figure 4.11d**). An intermolecular β-sheet is formed in both interfaces. However, unlike *Mm*EXTL2, the EXTL3 homodimer is stabilised by a pair of intermolecular disulphide bonds between Cys793 and Cys915. Nonetheless, these results suggest that different exostosins can homodimerise in a similar manner.

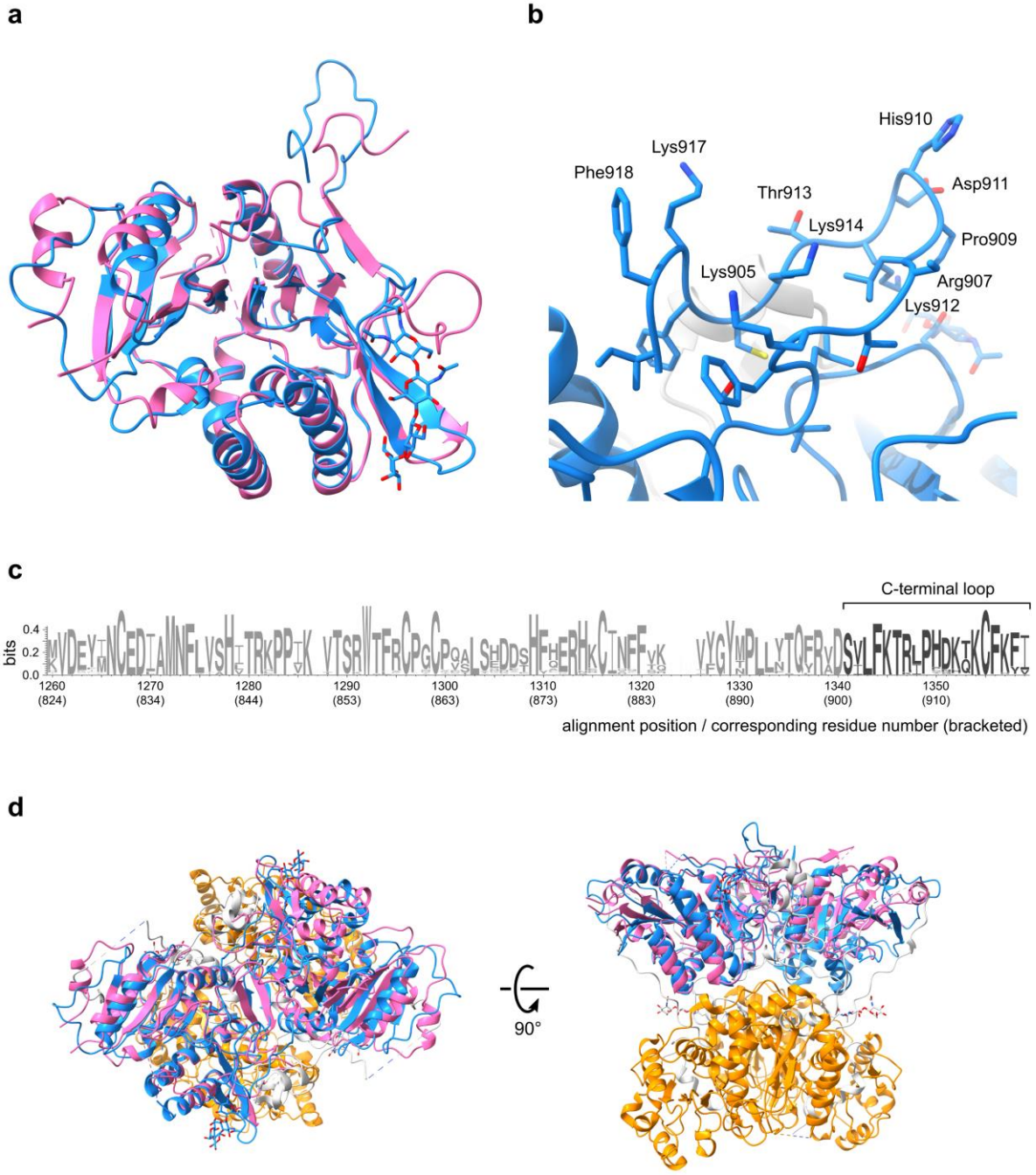


Figure 4.11 Structural features of the EXTL3 GT64 domain. a Monomeric GT64 domain of apo-EXTL3 (blue) aligned to that of mouse apo-EXTL2 (PDB: 1OMX; pink). An *N*-glycan attached to Asn790 of EXTL3 is shown in stick representation. **b** Close-up of the C-terminus of EXTL3, with sidechains displayed in stick representation. The sidechain of Arg907 was not fully modelled due to lack of density. **c** Sequence logo showing the C-terminal conservation (relative entropy in bits) between over 300 EXTL3 homologues from a wide range of metazoan species, created using WebLogo 3 (http://weblogo.threeplusone.com/). **d** Overlay of the full apo-EXTL3 structure with that of apo-EXTL2 (PDB: 1OMX; pink), aligned as in **a**.

4.2.6 The GT47 portion of EXTL3 constitutes an inactivated GT-B-fold domain

To confirm the fold of the GT47 domain, I reviewed its secondary structure. The domain appeared to be composed of two loosely connected α/β/α sandwich subdomains—consistent with the general architecture of the GT-B fold. Furthermore, when the isolated GT47 domain was submitted to the DALI server (Holm, 2020), which identifies similar structures from the Protein Data Bank (PDB), the top five best matches from the (reduced redundancy) PDB25 database were POGLUT1 from *D. melanogaster* (PDB: 5F84; RMSD = 3.9; GT90), T4 phage DNA β-glucosyltransferase (PDB: 1IXY; RMSD = 5.2; GT63), GtfC from *Streptococcus agalactiæ* (PDB: 4W6Q; RMSD = 4.0), PglH from *Campylobacter jejuni* (PDB: 6EJI; RMSD = 3.8; GT4), and T4 phage DNA α-glucosyltransferase (PDB: 1XV5; RMSD = 4.2; GT72)—all of which have been characterised as GT-B fold proteins. These results confirm that GT47-family glycosyltransferases adopt a GT-B fold. Nevertheless, the EXTL3 GT47 domain lacks some elements of secondary structure normally found in a typical GT-B domain: each subdomain contains only five β-strands, and across the protein merely eight substantial α-helices can be found (**Figure 4.12a–c**).

Although far from being a universal feature, a conserved aspartate/glutamate residue in the Cα4 helix—which typically participates in hydrogen bonding with the ribose moiety of the nucleotide sugar—is found in many GT-B-fold enzymes (Wrabl & Grishin, 2001; Hu & Walker, 2002; Breton *et al*, 2006; Martinez-Fleites *et al*, 2006). Furthermore, the phosphate moieties of the nucleotide sugar are typically bound at the N-terminus of this helix (Hu & Walker, 2002). Since I observed GlcAT-II activity from preparations of EXTL3ΔN, I investigated this helix to see whether it might also contain such features. Indeed, the EXTL3 Cα4 helix was found to contain a glutamate residue at the expected position (Glu453) (**Figure 4.12d**). However, surprisingly, this glutamate was found to form a salt bridge to the sidechain of Arg421, which might impede interaction with the nucleotide sugar. To investigate the conformation of this region further, I examined aligned structures from the top ten most similar GT families identified by DALI. Strikingly, not only were such salt bridges absent from the other enzymes, but it was also clear that the Cα4 helix of EXTL3 possesses one extra turn at its N-terminus compared with analogous helices in other GT-B structures. Moreover, it was apparent that this extension would most likely occlude the cavity that normally acts as a phosphate/donor sugar-binding pocket. These results suggest that the EXTL3 GT47 domain would have to undergo major conformational changes in order to bind UDP-GlcA or any other nucleotide sugar.

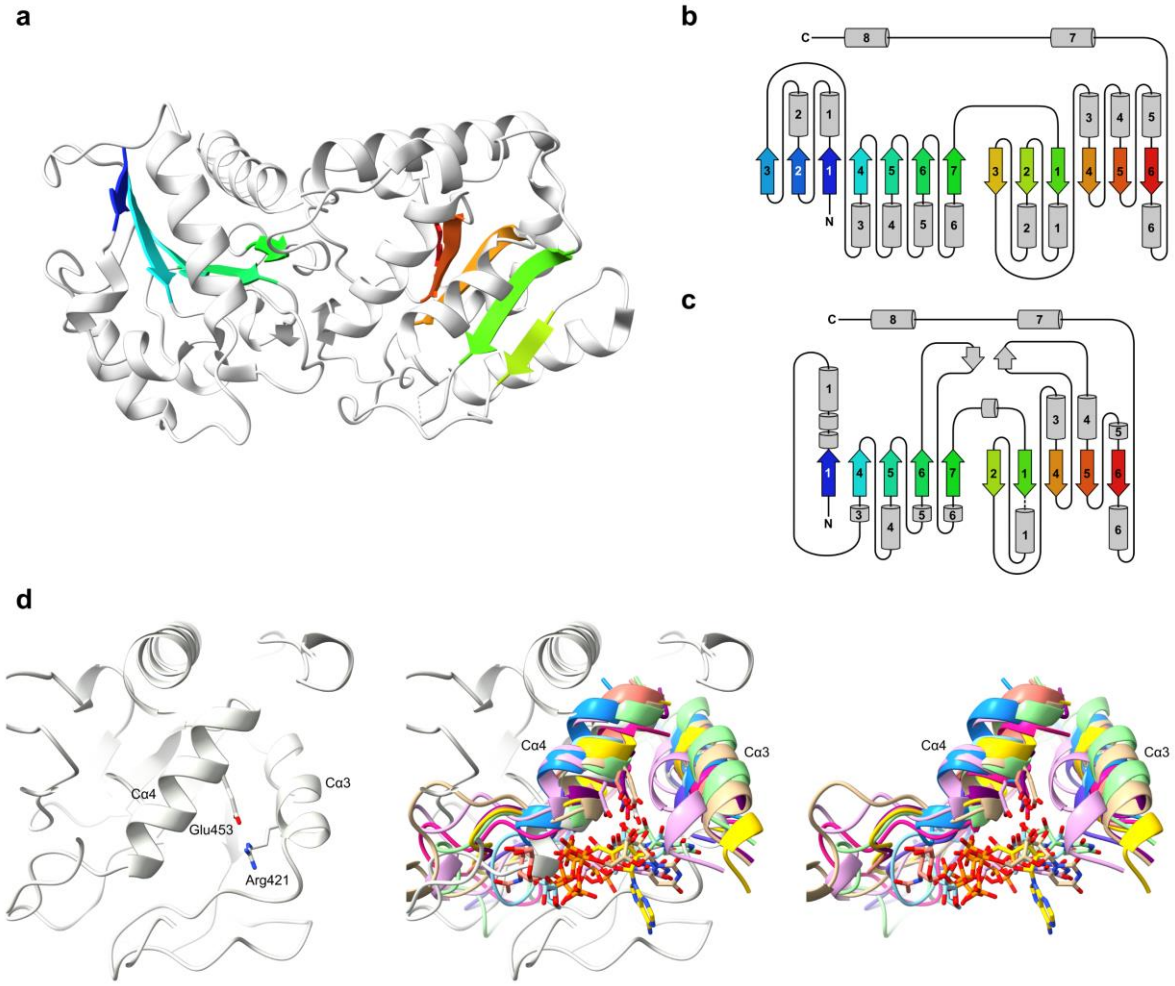

a
b
c
d
Cα4
Cα3
Glu453
Arg421
e
AtMUR3 PSSIMQMFQSSLFCL-QPQGD-----SY-TRRSAFDSMLAGCIPVFFHPG
AtXLT2 SSAILETFLGSDFCL-QPRGD-----SF-TRRSIFDCMLAGSIPVFFWRR
AtXUT1 SESVIELFRDSEFCL-QPPGD-----SP-TRKSIFDSLILGCIPVIFDPY
AtMBGT1 PVNVMKVFRNSVFCL-QPPGD-----SY-TRRSMFDSILAGCIPVFFHPG
AtXAPT1 PMTVLGVMARSRFCL-QAPGD-----SF-TRRSTFDAMLAGCIPVFFSPH
AtARAD1 RRAATKGMHTSKFCL-NPAGD-----TP-SACRLFDSIVSLCVPLIVSDS
AtXGD1 GKDYTKTMGMSKFCL-CPSGW-----EV-ASPREVEAIYAGCVPVIISDN
AtIRX10-L PTTYYEDMQRAIFCL-CPLGW-----AP-WSPRLVEAVIFGCIPVIIADD
AtExAD SDNYHKDIANSIFCG-AFPGD-----G--WSGRMEDSILQGCVPVIIQDG
HsEXT1 KYDYREMLHNATFCL-VPRGR-----RL-GSFRFLEALQAACVPVMLSNG
HsEXTL1 QTQRQETLPNATFCL-I-SGH-----RPEAASRFLQALQAGCIPVLLSPR
HsEXT2 VFDYPQVLQEATFCV-VLRGA-----RL-GQAVLSDVLQAGCVPVVIADS
HsEXTL3 ...REDRLELLKLSTFALIITPGDPRLVISSGCATRLFEALEVGAVPVVLGEQ
Cα3
Cα4

Figure 4.12 Structural features of the EXTL3 GT47 domain. a Isolated GT47 domain from apo-EXTL3, with central beta strands coloured from blue to red (from most N-terminal to most C-terminal. **b** Two-dimensional schematic of a typical GT-B topology. **c** Two-dimensional schematic of the EXTL3 GT47 topology. **d** The EXTL3 GT47 domain was aligned to 10 structurally similar GT-B enzymes using DALI (http://ekhidna2.biocenter.helsinki.fi/dali/). Panels show a close-up of the Cα3–Cα4 region in the C-terminal subdomain. For each structure, bound nucleotide/nucleotide sugar and the Cα4 Asp/Glu are shown if present. Left: EXTL3 alone. Centre: EXTL3 aligned to 5F84 (GT90; pale blue), 1IXY (GT63; pale pink), 4W6Q (pale green), 6EJI (GT4; salmon), 1XV5 (GT72; buff), 2NZW (GT10; hot pink), 6GNG (GT5; yellow), 1F0K (GT28; blue), 5UOF (GT20; burgundy), and 4BFC (GT30; purple). Right: DALI matches only. **e** Truncated sequence alignment of human and characterised Arabidopsis GT47 sequences, constructed using MUSCLE. Analogues to Glu453 and Arg421 in EXTL3 are shown in bold text.

These observations suggested that the EXTL3 GT47 domain is unusual in its conformation. To examine how widespread these features might be in the wider GT47 family, I aligned the EXTL3 GT47 domain sequence with those from the other human exostosins, as well as those from all characterised Arabidopsis GT47s. Interestingly, in all sequences excepting EXTL1, which is not thought to possess GlcAT activity (Kim *et al*, 2001), an Asp/Glu residue was aligned to the position of Glu453 in the EXTL3 sequence (**Figure 4.12e**). In contrast, an analogue for Arg421 was found only in EXTL1 and EXTL3, suggesting that an Arg421–Glu453-type salt bridge is not present in most GT47s. Furthermore, EXTL3 was alone in exhibiting a conspicuous insertion at the N-terminus of Cα4. Hence, other GT47 enzymes most likely possess a Cα4 helix with length similar to those in other GT-Bs. Together, these results suggest that EXTL3 is unique amongst the GT47 family in possessing its unusual active site features.

4.2.7 UDP binds to the GT64 domain of EXTL3, but not the GT47 domain

To investigate the ability of the EXTL3 GT domains to bind nucleotides, I solved the structure of EXTL3 a second time, this time in the presence of 10 mM UDP and 2.5 mM $MnCl_2$. Initial grid screening was performed by Tom Dendooven and Steven Hardwick, and, as before, Dima Chirgadze clipped the grids and operated the microscope; however, for this structure, I carried out grid freezing and all of the data processing. I produced a density map at 2.9 Å resolution (**Figure 4.13**), using which I was able to refine a new co-ordinate map based on the structure

of the apo-enzyme (**Figure 4.14**, **Figure 4.15**). Firstly, I inspected density belonging to the GT64 domain. New density corresponding to UDP and an Mn^{2+} ion was clearly visible in the predicted nucleotide binding pocket. As expected, the Mn^{2+} ion is co-ordinated by the sidechain of Asp746 in the DxD motif, as well as the phosphate moieties of UDP. The ribose moiety of UDP itself is bound through hydrogen bonding to the sidechain of Asp745 and the main-chain carbonyl group of Leu668; the uracil base is bound via hydrogen bonds to Asn697 and Asn723, as well as a parallel-displaced π-stacking interaction with the sidechain of Tyr670. These protein-ligand interactions are highly similar to those seen in the previously reported structure of *Mm*EXTL2 bound to UDP-GlcNAc (PDB:1ON6; **Figure 4.16**).

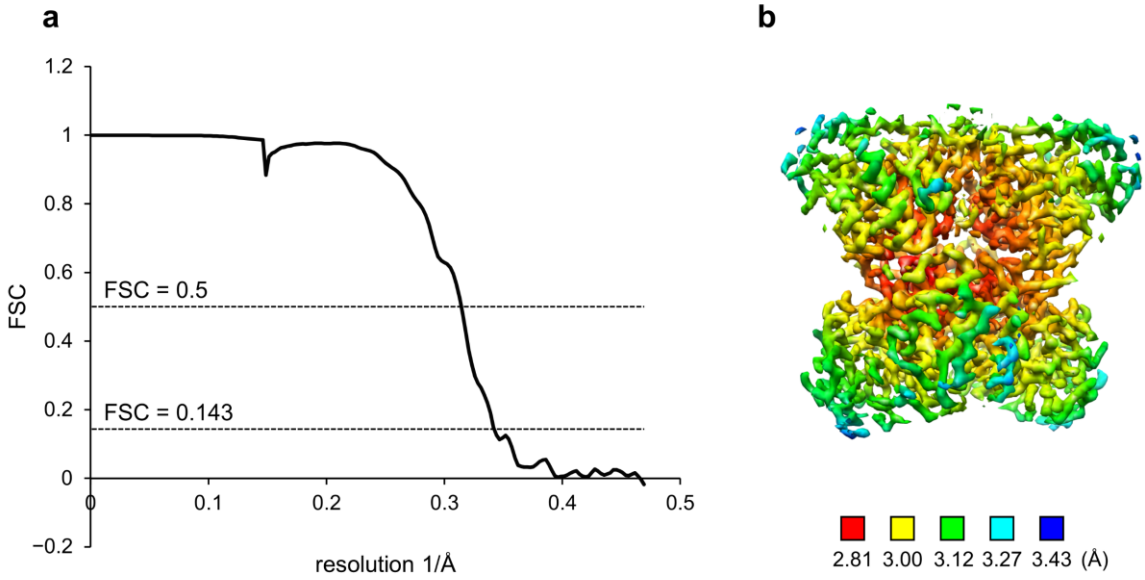


Figure 4.13 Resolution estimates for the UDP-bound EXTL3 density map. a FSC curve. **b** Local resolution map, calculated using RELION.

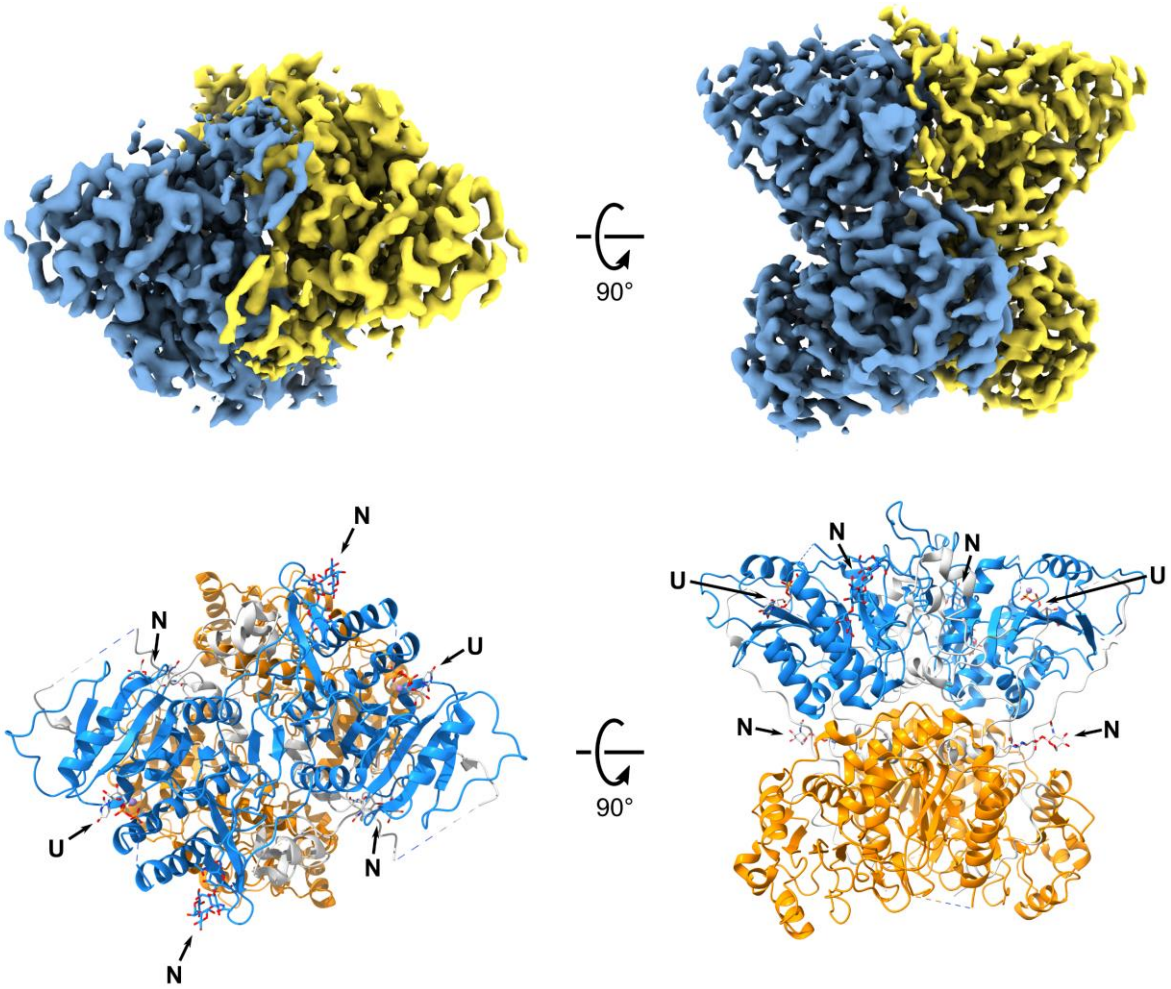


Figure 4.14 Overall structure of the EXTL3 catalytic domain dimer bound to UDP. a Map and model. Top: cryo-EM density map at contour level 0.027 (top and side views). The two constituent monomers are shown in blue and yellow, respectively. Bottom: corresponding atomic model. GT47 domains are shown in orange; GT64 domains are shown in blue. UDP and *N*-glycans attached at Asn592 and Asn790 are shown in stick representation and labelled as **U** and **N**, respectively.

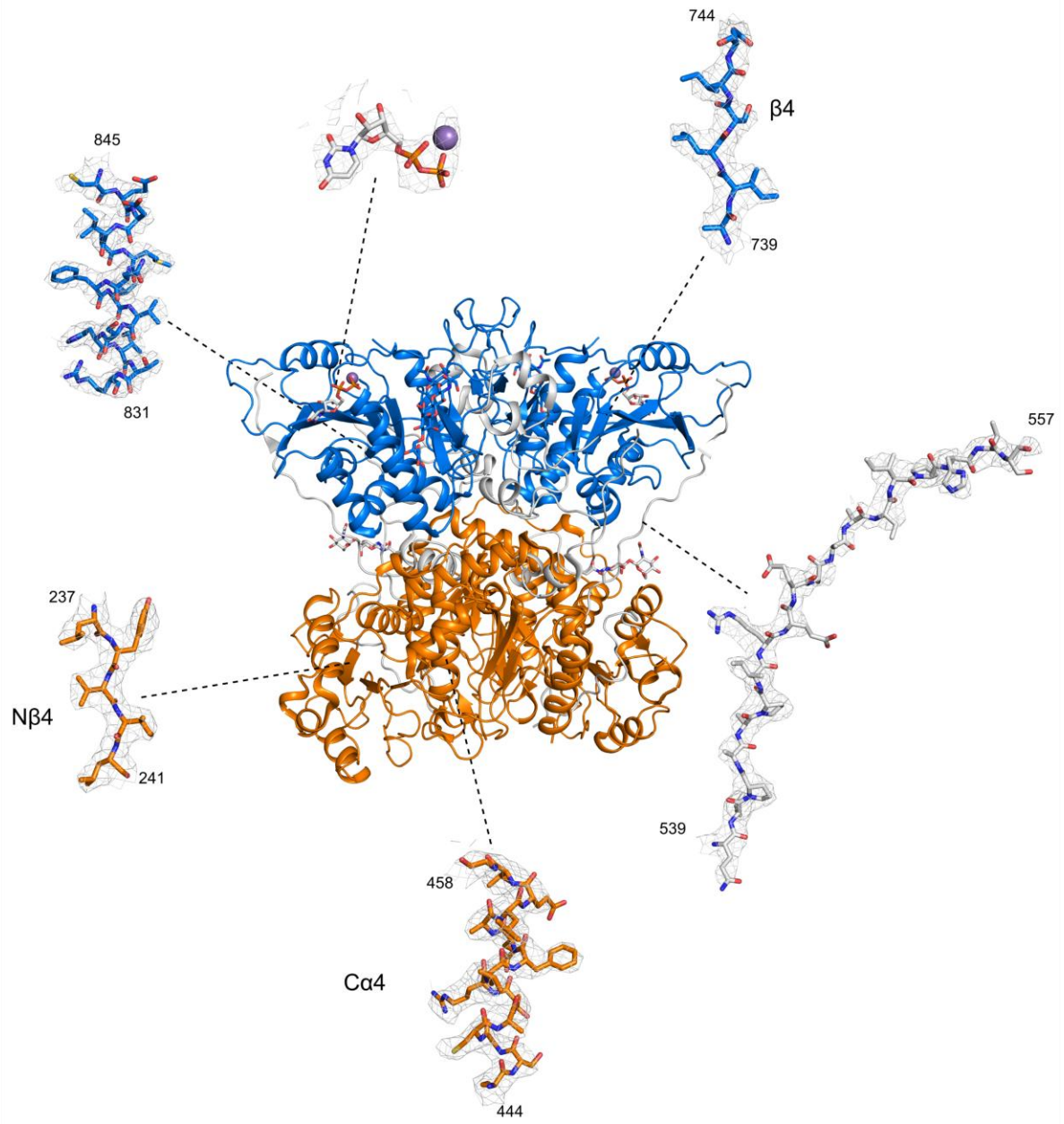


Figure 4.15 Local cryo-EM density at key areas of the UDP-bound EXTL3 map.

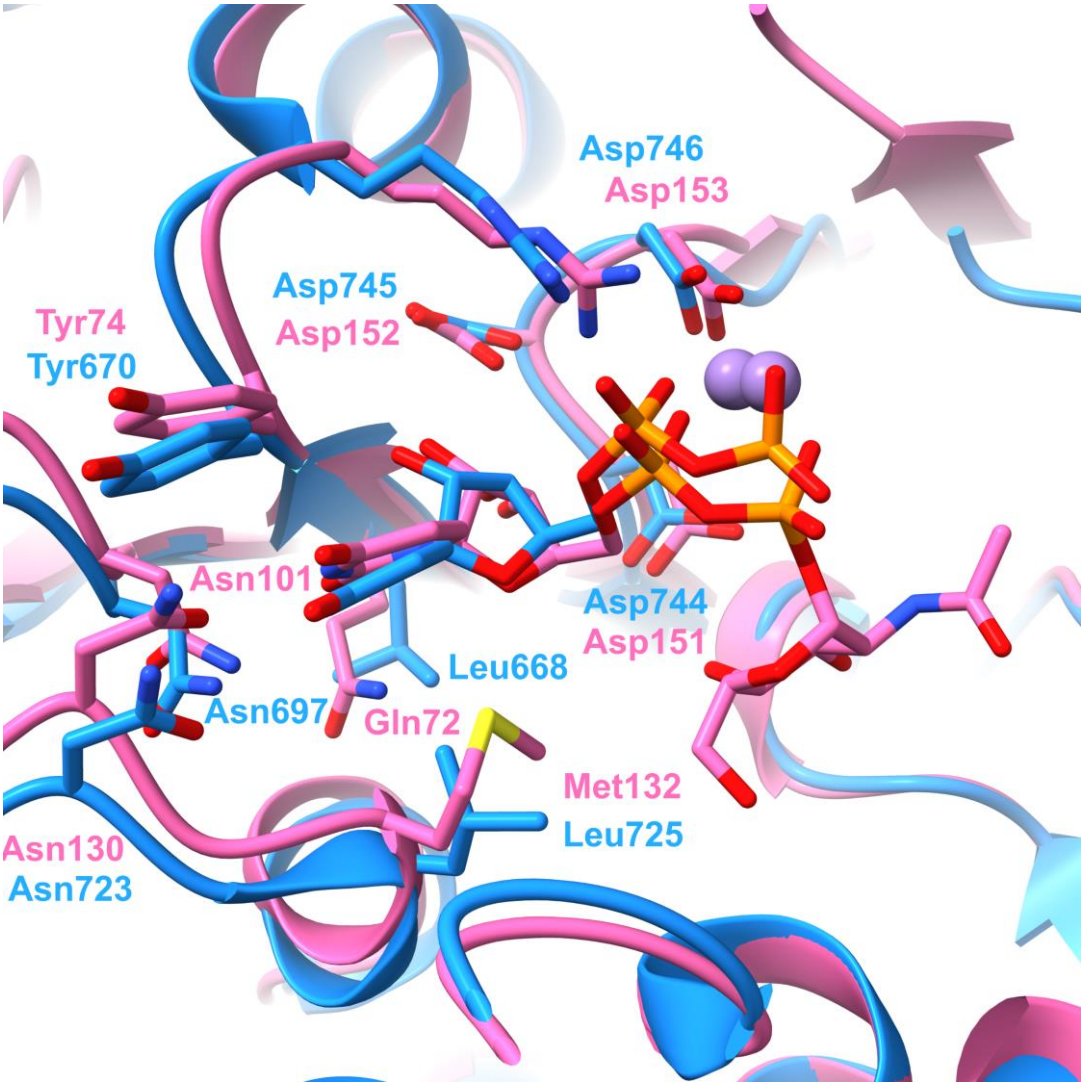


Figure 4.16 Nucleotide-binding residues are conserved between human EXTL3 and mouse EXTL2. GT64 domain structural alignment of UDP-bound EXTL3 (blue) and UDP-GlcNAc-bound EXTL2 from mouse (PDB: 1ON6; pink) (close-up of nucleotide binding site). For EXTL3, bound UDP and Mn^{2+} are shown as blue/heteroatom-coloured sticks and a purple sphere, respectively. Similarly, for EXTL2, bound UDP-GlcNAc and Mn^{2+} are shown in pink/heteroatom-coloured sticks and a purple sphere, respectively.

I also inspected the portion of the density map corresponding to the GT47 domain. Despite the high concentration of UDP used, no discernible new density was present, however. This result suggests that the GT47 domain is unable to bind UDP—and is therefore also unlikely to bind UDP-GlcA.

Nevertheless, I attempted to use the EXTL3 GT47 structure to predict residues involved in UDP-GlcA binding in EXT1, which has previously been demonstrated to possess GlcAT-II activity *in vitro* (Busse-Wicher *et al*, 2014). Indeed, experiments by Wei *et al* (2000) on the EXT1 orthologue from Chinese hamster (*Cg*EXT1) have previously demonstrated that GlcAT-II activity originates from the GT47 domain of this protein. In this study, six point mutations were identified and shown to abrogate *Cg*EXT1 GlcAT-II activity. I aligned the protein sequence of *Cg*EXT1 to the sequences of human exostosins in order to determine how these mutations might relate to the EXTL3 protein structure. Interestingly, four of the six mutations mapped to positions just preceding or within the predicted Cα4 helix (**Figure 4.17**). Interestingly, the site of one of these mutations, E349K, was aligned to Glu453 in EXTL3. These results support the notion that, like in other GT-B enzymes, the nucleotide sugar is bound in close proximity to the Cα4 helix in Chinese hamster EXT1, as well as in its close homologue in humans.

The physical association of GT47 and GT64 domains within exostosins has been proposed to allow substrate channeling between the two active sites (Young, 2004). Hence, it might be expected that the two active sites would be positioned close together, or that a mechanism to shuttle substrates between the two domains might exist. I postulated that the distance between the active sites in EXT1 is very likely to be similar to the distance between the analogous sites in EXTL3. Therefore, to estimate their proximity, I inspected the distance between the Cα4 helix of the GT47 domain and UDP bound in the GT64 domain of EXTL3. I calculated the gap between these two sites to measure over 45 Å (**Figure 4.18**). Hence, it seems unlikely that substrates can be mechanistically shuttled between the two active sites in EXT1.

Figure 4.17 GlcAT-II-inactivating mutations in *Cg*EXT1 can be mapped to the GT47 domain of EXTL3. Sequence alignment of EXT1 from Chinese hamster (*Cg*EXT1) to EXT1, EXT2, and EXTL3 from *Homo sapiens*, truncated to the GT47 domain. The secondary structure of EXTL3 is annotated above the sequences (H = α-helix; E = β-sheet). The locations of point mutations that abrogate *Cg*EXT1 GlcAT-II activity, identified by Wei *et al* (2000), are highlighted in black boxes. Arg340 in EXT1 and Arg223 and Asp227 in EXT2, whose mutation can cause hereditary multiple exostoses, are highlighted in grey.

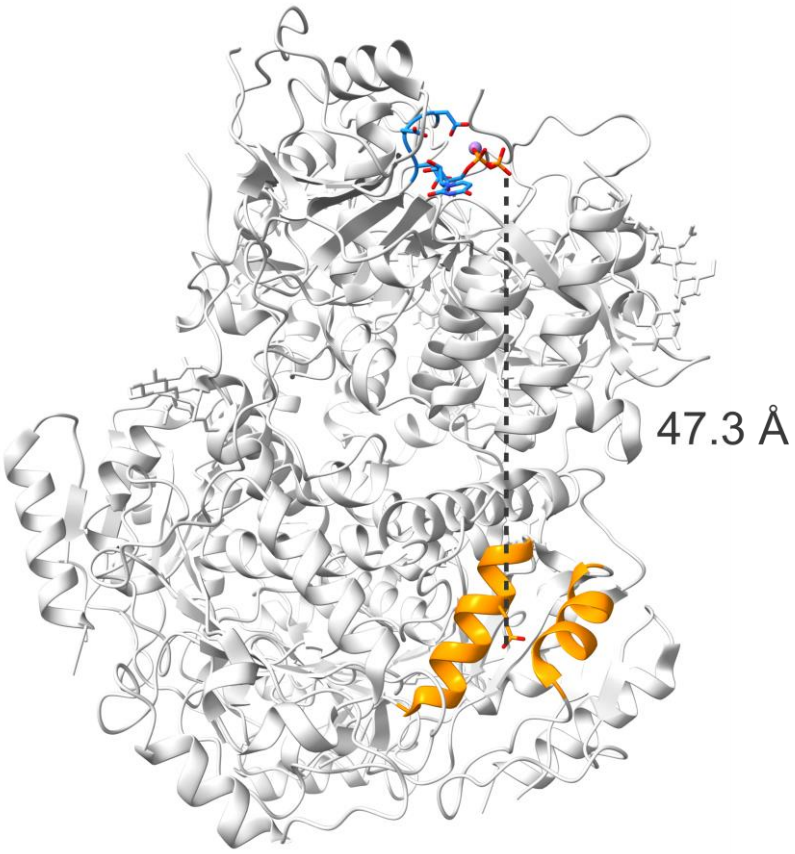


Figure 4.18 The GT64 domain active site and the inactivated GT47 domain active site are distant from one another in EXTL3. The DxD motif and bound UDP molecule of the GT64 domain are shown in blue/heteroatom colouring, whereas the Cα3 and Cα4 helices of the GT47 domain are shown in orange. The closest distance between Glu453 and the β-phosphate of UDP was measured using PyMOL.

4.2.8 The EXTL3 structure reveals how some pathogenic missense mutations disrupt EXTL3 function

Several mutations to the *EXTL3* gene are known to result in developmental and neurological pathologies in humans (Guo *et al*, 2017; Volpi *et al*, 2017; Oud *et al*, 2017). Of these, six missense mutations have been characterised, and relate to the amino acid changes P318L, R339W, P461L, R513C, N657S, and Y670D (Yamada, 2020). I reasoned that the EXTL3 structure might be able to explain the effects of these mutations at the molecular level. Accordingly, I attempted to determine the roles of these residues in EXTL3 function by finding their position within the overall protein structure. Interestingly, I found that the majority are not localised to the GT64 domain active site; instead, many of the residues appear to be involved in linking together different elements of the secondary structure together within the

core of the enzyme (**Figure 4.19**). For example, the penultimate α-helix of the GT47 domain is linked to the C-terminus of the last helix of the GT64 domain by virtue of a hydrogen bond between the sidechain of Arg513 in the former and the main-chain carbonyl group of Val886 in the latter. Hence, the characterised R513C mutation may reduce the association between the GT47 and GT64 domains, or potentially introduce an inappropriate disulphide bridge to a

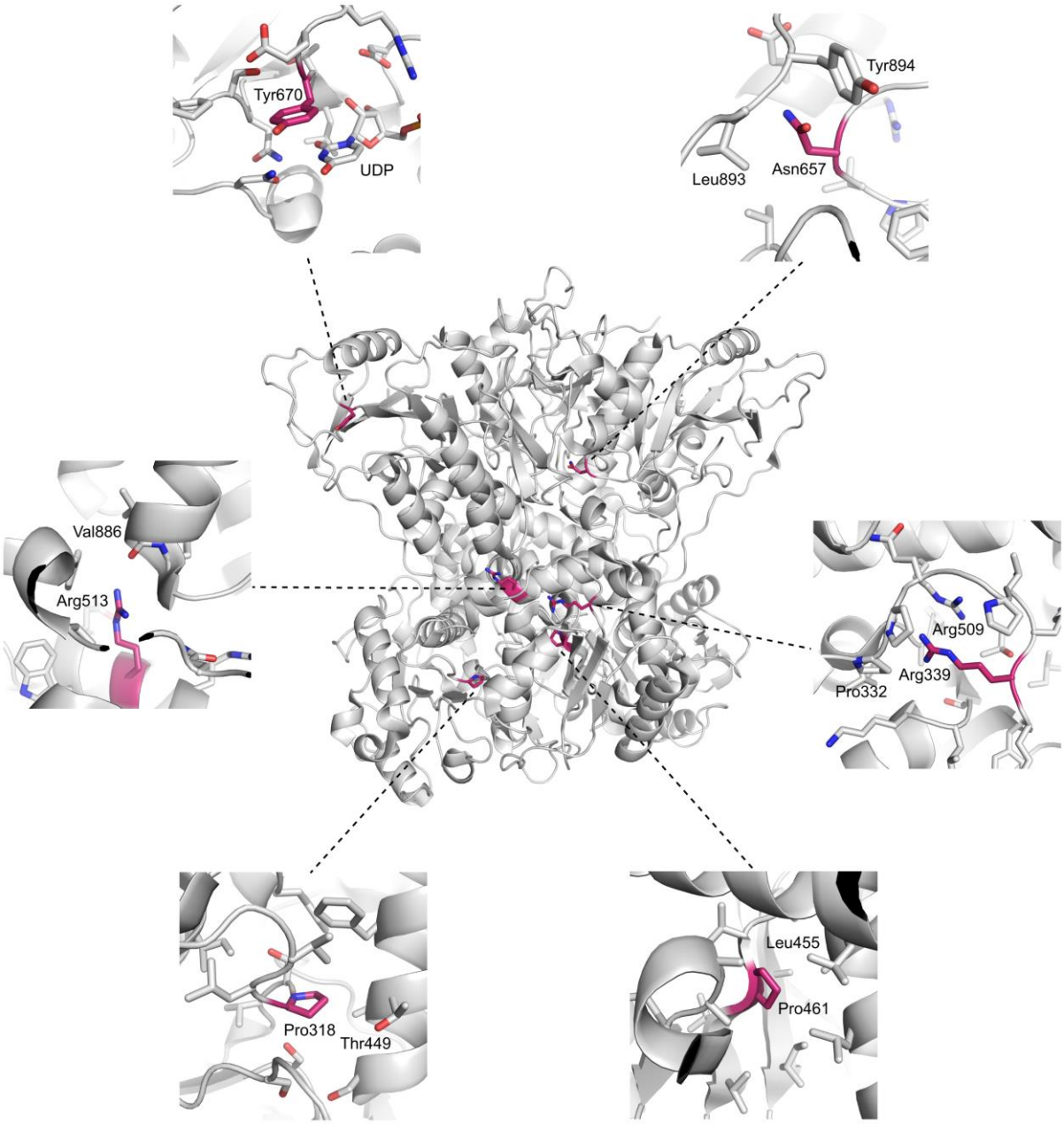


Figure 4.19 Pathogenic missense mutations in the human *EXTL3* gene can be mapped onto the EXTL3 structure. Residues whose mutation has been shown to cause disease in human patients are shown in purple (within their structural context). Close-ups, from top left (clockwise): Tyr670, Asn657, Arg339, Pro461, Pro318, and Arg513.

cysteine elsewhere in the protein—either way, preventing correct folding. Indeed, fibroblasts from individuals with this mutation are reported not to exhibit detectable EXTL3 protein in their Golgi apparatus, leading to the suggestion that EXTL3[R513C] is mislocalised or degraded in these cells (Oud *et al*, 2017). In contrast, I showed above that Y670 is involved in binding the nucleotide sugar in the GT64 domain active site. Hence, the Y670D mutation is likely to reduce GlcNAcT activity without necessarily affecting the expression or stability of the EXTL3 protein. This is in accordance with the reported result that HS levels are reduced in the serum and urine of individuals with this mutation, but that EXTL3[Y670] can still be detected in fibroblast Golgi bodies (Oud *et al*, 2017).

I also attempted to explain some pathogenic missense mutations in *EXT1* and *EXT2* by virtue of their homology to *EXTL3*. Deleterious mutations to these genes cause hereditary multiple exostoses (HME) (Zak *et al*, 2002). For instance, one of the most common pathogenic amino acid changes observed in EXT1 involves the mutation of Arg340 (R340H, R340S, R340C, and R340L mutations have been reported) (Wuyts *et al*, 1998; Ishimaru *et al*, 2016; Fusco *et al*, 2019). By consulting my alignment of exostosin protein sequences (**Figure 4.17**), I found that this residue is likely located on the loop situated at the N-terminus of Cα4, which, as discussed above, is likely proximal to bound donor substrate. Indeed, cationic residues in this loop are known to aid the binding of phosphate moieties in other GT-Bs (Albesa-Jové *et al*, 2014)—the function of Arg340 may lie in this role, therefore. Hence, these mutants may be unable to bind UDP-GlcA. Similarly, I was able to identify the likely roles of Arg223 and Asp227 in EXT2, which are frequently affected by R223P and E227N mutations in patients with HME, respectively (Shi *et al*, 2000; Dobson-Stone *et al*, 2000; Gentile *et al*, 2019; Ishimaru *et al*, 2016; Fusco *et al*, 2019). Based on my alignment, these residues likely correspond to Arg309 and Asp313 in EXTL3, which form a salt bridge within the N-terminal half of the GT47 domain. Hence, mutations in these residues may disrupt the folding of this domain.

4.2.9 Lower plant genomes exhibit GT47 sequences closely related to exostosins

As discussed in *Section 1.4.2*, typical plant genomes encode many more GT47 enzymes than animal genomes do (Geshi *et al*, 2011). These plant enzymes have been classified into the six GT47-A–F clades (Li *et al*, 2004). However, where animal GT47 domains stand in relation to these clades is currently unknown. With the structure of EXTL3 available, however, it is now possible to make informed sequence truncations in order to construct meaningful alignments between animal and plant GT47 protein sequences. Hence, I downloaded a GT47-family hidden Markov model (HMM) from the dbCAN2 server (Yin *et al*, 2012; Zhang *et al*, 2018a)

and used it to search for GT47 protein sequences from proteome models of *Homo sapiens*, *Drosophila melanogaster*, *Cænorhabditis elegans*, *Amphimedon queenslandica* (a sponge), and *Monosiga brevicollis* MX1 (a choanoflagellate), as well as *Arabidopsis thaliana*, *Ginkgo biloba* (a gymnosperm 'living fossil'), and *Physcomitrium patens* (a moss). As expected, the sequences extracted from the four metazoan species all contained GT64 domains (with the exception of rib-1 from *C. elegans*). I was also able to detect GT47 domain-containing sequences from *M. brevicollis* MX1, including one sequence with a predicted N-terminal sulphatase domain. After crudely removing GT64 and sulphatase domains, all sequences were aligned before careful truncation to the GT47 domain portion (corresponding to residues 176–538 of EXTL3). I then constructed a phylogeny using RAxML (Stamatakis, 2006, 2014). The results revealed that, as seen previously, the Arabidopsis sequences could be grouped into six clades (**Figure 4.20**). However, surprisingly, for *G. biloba* and *P. patens* sequences, a seventh well supported clade was apparent; furthermore, metazoan and choanoflagellate sequences were also grouped in this clade. As previously reported (Feta *et al*, 2009), the metazoan sequences were themselves split into three major groups corresponding to homologues of EXT1, EXT2, and EXTL3, respectively. The branch lengths and apparent topology suggested that the *G. biloba* and *P. patens* sequences within this clade are closely related to these animal sequences. Hence, I propose that this clade is named GT47-G. Furthermore, *M. brevicollis* appeared to exhibit several sequences that were not grouped in GT47-G, instead being distributed throughout the tree. These results suggest that the various clades of GT47 enzymes observed in plants likely diverged at an early point in eukaryotic evolution.

4.2.10 The difference between UDP-sugar-binding and GDP-sugar-binding GT64s is reflected in their protein sequences

Although the evolutionary relationships between animal and plant GT64 sequences have been studied previously (Edvardsson *et al*, 2011), I also created a phylogeny of GT64 sequences from these species for completeness. Accordingly, I searched for sequences using a GT64 HMM from dbCAN2. The sequences were truncated to their GT64 domain after alignment. Remarkably, one of the hits from *P. patens* (Pp3c16_2260V3.1, later determined to be the only *P. patens* homologue of uncharacterised Arabidopsis protein AT1G80290) was found to be 932 amino acids in length. Secondary structure prediction and distant homology searches using Jpred 4 (Drozdetskiy *et al*, 2015) and HHpred (Zimmermann *et al*, 2018; Gabler *et al*, 2020) revealed that, in addition to its N-terminal GT64 domain, this protein likely contains two GT8-related GT-A domains. Such a domain structure is interesting because both characterised

Arabidopsis GT64 enzymes (GINT1 and GMT1) are known to act on the product of a GT8 enzyme (IPUT1). Hence, this tri-domain glycosyltransferase may have evolved to increase the efficiency of a related glycosylation pathway.

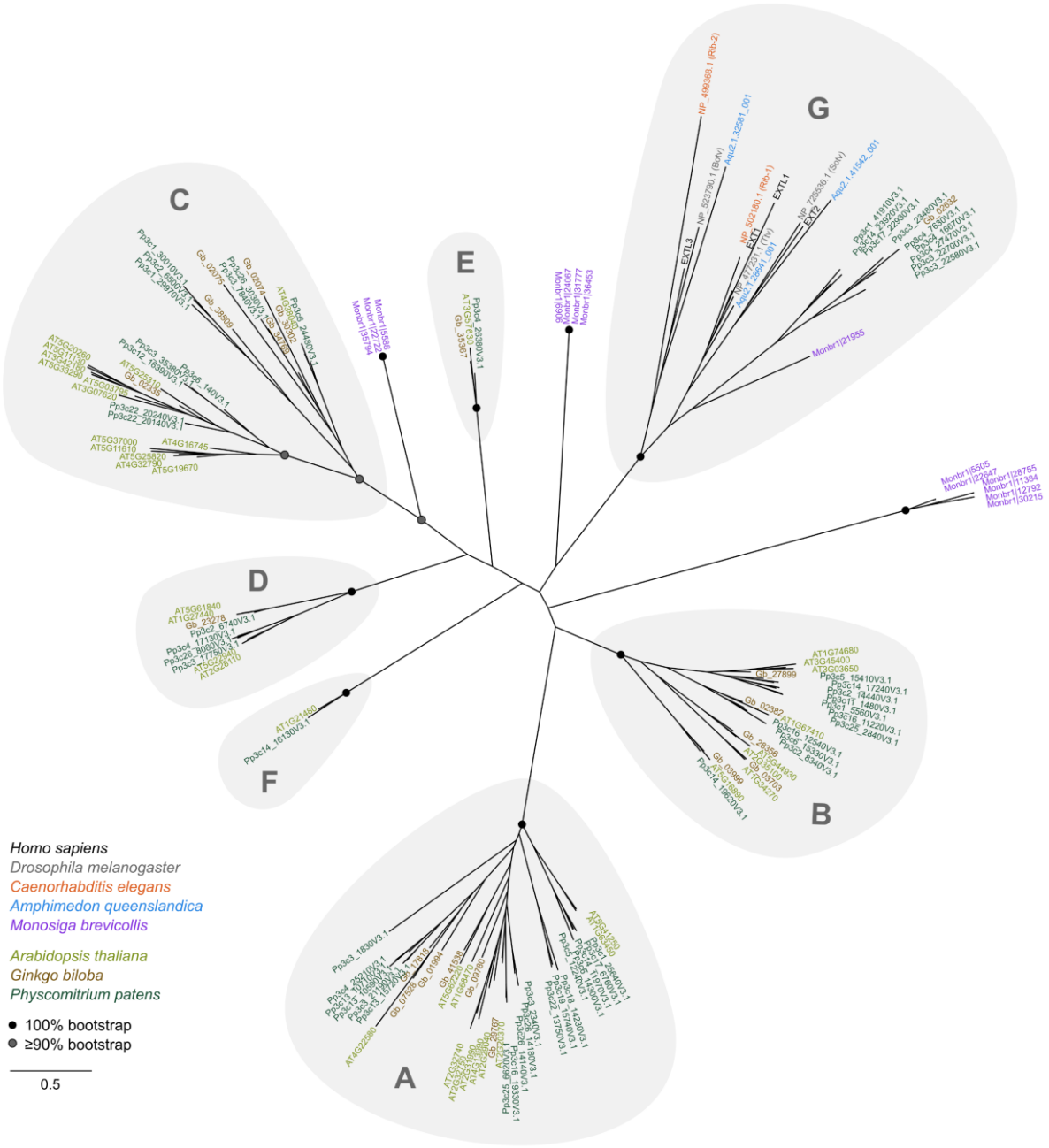


Figure 4.20 Phylogeny of GT47 sequences from animals, *Monosiga brevicollis*, and plants. A GT47 HMM from dbCAN2 (http://bcb.unl.edu/dbCAN2/) to identify GT47 domain-containing sequences from the indicated species. Sequences were aligned before truncation to their GT47 domain, and a phylogeny was constructed using RAxML with 100 rapid bootstraps. Bootstrap values above 90% are indicated for major branching points.

Subsequently, I constructed a phylogeny using RAxML. In the resultant tree, the plant enzymes were grouped into three well supported groups, each containing one of the three known Arabidopsis GT64 enzymes (**Figure 4.21a**). In contrast to the GT47 tree, no animal sequences were seen to be grouped in any of the clades containing plant sequences. As reported previously (Edvardsson *et al*, 2011), EXTL2 appeared to constitute the earliest-diverging animal GT64 sequence. However, the splits forming the base of the tree were not well supported; hence, strong conclusions could not be drawn about the evolutionary origins of the different clades.

From a structural biology perspective, the evolution of the GT64 family is of particular interest because it contains not only enzymes that use UDP-sugars (UDP-GlcNAc by exostosins and GINT1) but also enzymes that use GDP-sugars (GDP-Man by GMT1). I speculated that the residues involved in base binding might differ between these enzymes. Hence, I re-examined the aligned sequences with reference to the EXTL3 and *Mm*EXTL2 structures. As shown above in **Figure 4.16**, three uracil-binding residues were found to be identical in both structures: Tyr670/Tyr74 (at the C-terminus of β1), Asn697/Asn101 (at the C-terminus of β2), and Asn723/Asn130 (at the N-terminus of α3). Whereas the tyrosine participates in a stacking interaction, the sidechains of both asparagine residues appear to make Watson-Crick-esque hydrogen bonds to the uracil ring. Interestingly, the aligned amino acids from GINT-related sequences (which presumably all encode α-*N*-acetylglucosaminyltransferases) were also identical to these residues (**Figure 4.21b**). In contrast, however, GMT-related sequences (which presumably all encode α-mannosyltransferases) exhibited different residues at all three sites. Interestingly, tyrosine was replaced by tryptophan in these sequences. It is tempting to conclude that the bicylic sidechain of this tryptophan residue may help to selectively bind the bicylic structure of the guanine base through a more favourable stacking interaction. Similarly, the replacement of the two asparagine residues with serine and aspartate may permit the formation of appropriate hydrogen bonds to the base. The third group of plant GT64s, which includes the uncharacterised AT1G80290, appear to exhibit Phe/Tyr at the C-terminus of β1 rather than tryptophan, suggesting that they might bind a UDP-sugar; however, since the two UDP-binding asparagine residues were not conserved in these sequences, a strong prediction cannot be made about their nucleotide specificity.

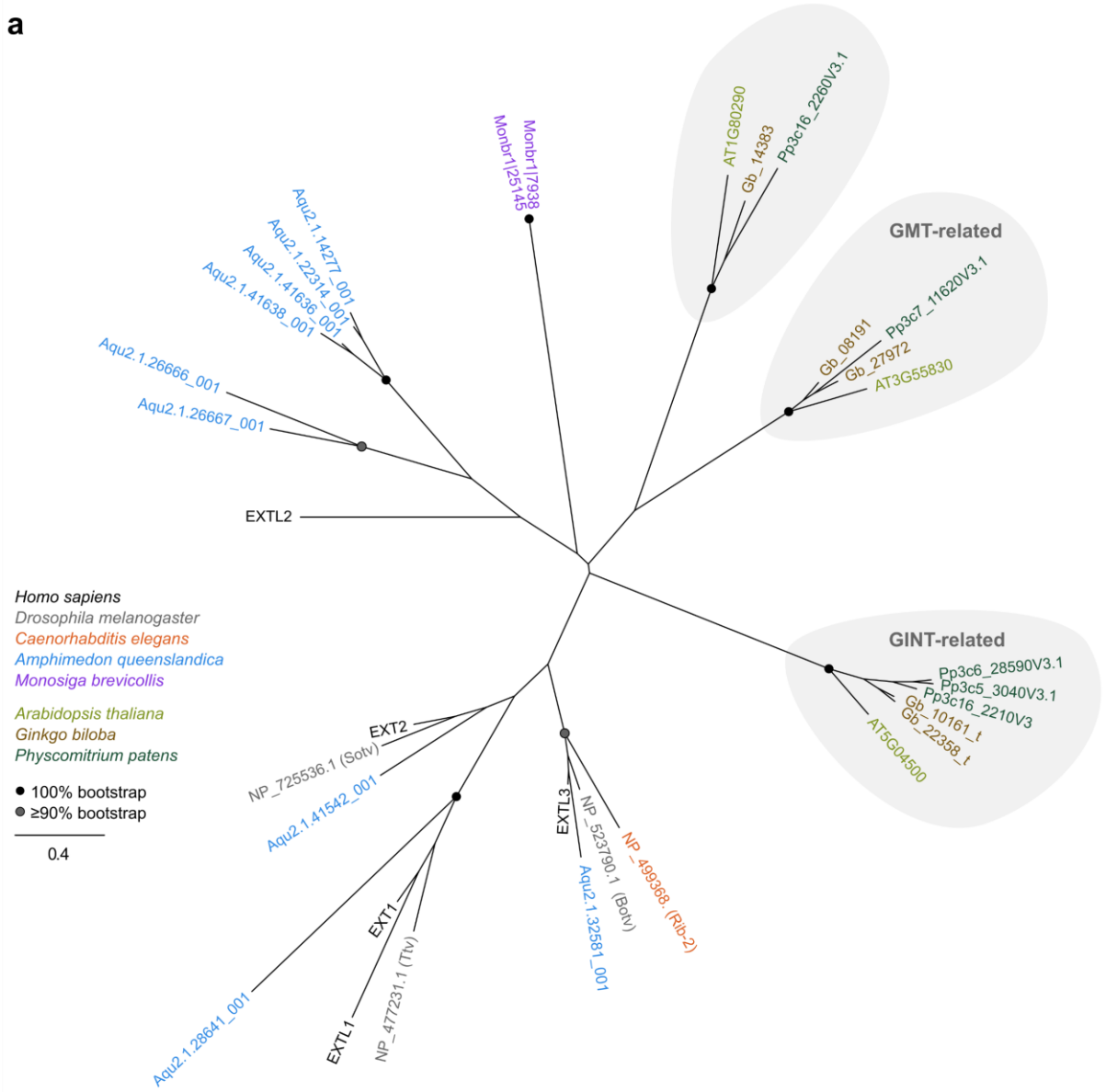

a
GMT-related
GINT-related
AT1G80290
Gb_14383
Pp3c16_2260V3.1
Monbr1|7938
Monbr1|25143
Aqu2.1.14277_001
Aqu2.1.22314_001
Aqu2.1.41636_001
Aqu2.1.41638_001
Aqu2.1.26666_001
Aqu2.1.26667_001
EXTL2
Gb_08191
Pp3c7_11620V3.1
Gb_27972
AT3G55830
Pp3c6_28590V3.1
Pp3c5_3040V3.1
Pp3c16_2210V3
Gb_10161_t
Gb_22358_t
AT5G04500
EXT2
NP_725536.1 (Sotv)
Aqu2.1.41542_001
EXTL3
NP_523790.1 (Botv)
NP_499368 (Rlb-2)
Aqu2.1.32581_001
EXT1
NP_477231.1 (Tlv)
EXTL1
Aqu2.1.28641_001
Homo sapiens
Drosophila melanogaster
Caenorhabditis elegans
Amphimedon queenslandica
Monosiga brevicollis
Arabidopsis thaliana
Ginkgo biloba
Physcomitrium patens
100% bootstrap
≥90% bootstrap
0.4
b
β1
Y670
α1
β2
N697
β3
N723
α3
EEEEEEE--------HHHHHHHHHHH------EEEEEEEE----------EEEEE----HHHHHHH
EXTL3 FTVVMLT----YE--REEVLMNSLERLNGLPYLNKVVVVWNSP-KLPSEDL-LWPD-IG----------VPIMVVRTEKNSLNNRFL
EXT1 FTAVIHAVTPLVS--QSQPVLKLLVAAAKSQYCAQIIVLWNCD-KPLPAKH-RWPA-TA----------VPVVVIEGESKVMSSRFL
EXTL1 FSALI------WVGPPGQPPLKLIQAVAGSQHCAQILVLWS---NERPLPS-RWPE-TA----------VPLTVIDGHR-KVSDRFY
EXT2 FTAIVLT----YD--RVESLFRVITEVSKVPSLSKLLVVWNNQNKNPPEDS-LWPK-IR----------VPLKVVRTAENKLSNRFF
EXTL2 FTLIMQT----YN--RTDLLLKLLNHYQAVPNLHKVIVVWNNI-GEKAPDE-LWNS-LGPHP-------IPVIFKQQTANRMRNRLQ
GINT
Pp3c16_2210V3.1 FTMIAMT----YEA-RLWNLQMYVKHYSRCASVREIVVVWNK--GTPPD-L-EDFD-SA----------VPVRIRVEPQNSLNNRFK
Pp3c6_28590V3.1 FTMVAMT----YDA-RLWNLQMYVKHYSRCTSVREIVVVWNK--GTPPNPA-LDFD-SA----------VPVRIRVEPKNSLNNRFK
Pp3c5_3040V3.1 FTMIAMT----YDA-RLWNLQLYVKHYSRCASVREIVVVWNK--GIPPDPV-LDFD-SA----------VPIRIRVEPENSLNNRFK
Gb_10161 FTMLTMT----YEA-RLWNLKMYIKHYSRCASVREIVVVWNK--GQPPDPK-TDFD-ST----------VPVRIRVEEQNSLNNRFKF
Gb_22358 FTLLTMT----YEA-RLWNLKMYIKHYSRCASVREIVVVWNK--GQPPNPE-TDFD-ST----------VPVRIRVEKQNSLNNRFK
AT5G04500 FTLATMT----YDA-RLWNLKMYVKRYSRCPSVKEIVVIWN---KGPPPDL-SELD-SA----------VPVRIRVQKQNSLNNRFE
GMT
Pp3c7_11620V3.1 YTVVINT----WK--RNDLLKRSVSHYSSCQGVDAIRVVWSEPTPPSDSLR-SSLEGLVELATRKKHRHVSLQLDIHVDDDLNNRFK
Gb_08191 YTVLINT----WK--RNDLLKQSVAHYAACNSVDAIHVVWSETDSPSDSLQ-AYLK-RIVQLKSQRTKKAEFRFDLNEVDNLNNRFR
Gb_27972 YTLLINT----WK--RNSLLKQAVAHYACCSSVDAIRVVWSENDPPSESLR-TYLR-KAVHSKSKSINKPDLRFDLNEEDNLNNRFK
AT3G55830 YTLLMNT----WK--RYDLLKKSVSHYASCSRLDSIHIVWSEPNPPSESLK-EYLH-NVLKKKTRDGHEVELRFDINKEDSLNNRFK
Pp3c16_2260V3.1 LTILVNG----FGEARLPLLEASVRKYSSSPVVHSVFVLWGNTSTPDSFLQASKFQSIG----------APIYIVRQNSMSLNDRFL
Gb_14383 LTVLING----FSETRLALLKKITRTYSASPCVHSIFILWGNNSTPSETLE-KQVF-DSLG--------APIYIIKQHTASLNNRFL
AT1G80290 ITVLING----YSEYRIPLLQTIVASYSSSSIVSSILVLWGNPSTPDQLLD-QLYQNLTQYSPGS----ASISLIQQSSSSLNARFL

Figure 4.21 Phylogeny of GT64 sequences from animals, *Monosiga brevicollis*, and plants. **a** A GT64 HMM from dbCAN2 (http://bcb.unl.edu/dbCAN2/) was used to identify GT64 domain-containing sequences from the indicated species. Sequences were aligned before truncation to their GT64 domain. Several sequences from *Amphimedon queenslandica* were removed from the alignment due to their short length. A phylogeny was subsequently constructed using RAxML with 100 rapid bootstraps. Bootstrap values above 90% are indicated for major branching points. **b** Alignment of GT64 domain sequences from the human exostosins and plant GT64 enzymes, truncated to the nucleotide-binding part of the GT-A fold (β1–α3) for brevity. The secondary structure of EXTL3 is annotated above the sequences (H = α-helix, E = β-sheet). The three uracil-binding residues in EXTL3 found to be present in mouse EXTL2 are labelled (Tyr670, Asn697, and Asn723); aligned analogues are highlighted in bold text.

4.3 Discussion

A wide range of sulphated polysaccharides can be found in eukaryote species (Aquino *et al*, 2011; Jiao *et al*, 2011; Yamada *et al*, 2011; Mourão *et al*, 2018). However, the most intensely studied of these have been sulphated glycosaminoglycans in animals. In particular, HS and CS stand out as the most highly conserved and functionally important of these extracellular glycans throughout animal evolution.

My bioinformatic results shed some surprising insights into the evolution of HS backbone synthesis enzymes. For example, I was able to identify sequences in the genomes of *P. patens* and *G. biloba* that appeared to exhibit a close relationship with animal GT47 sequences; I named the clade that contains these sequences 'GT47-G'. Curiously, sequences from this clade were not included in a previous phylogeny of *P. patens* GT47 enzymes (Geshi *et al*, 2011), perhaps due to the lack of a direct homologue in Arabidopsis, or perhaps due to the evolutionary distance between these sequences and those in other GT47 clades. Nevertheless, the presence of GT47-G enzymes in such a wide range of eukaryotes suggests that the founder of this subfamily belonged to an early eukaryote. Furthermore, the existence of this cosmopolitan clade, as well as the existence of choanoflagellate GT47s spread throughout the GT47 tree, suggests that all *seven* of these subgroups may have emerged at a very early stage in eukaryotic evolution. This hypothesis is in conflict with several previous hypotheses that proposed that all plant GT47s are descended from algal GT47-Es (Geshi *et al*, 2011; Ulvskov *et al*, 2013; Møller *et al*, 2017; Xu *et al*, 2018).

The point at which GT47-G enzymes obtained a HS-synthesising activity is less clear, however. For a start, one might suppose that the fusion of two early GT47 and GT64 proteins to form the first bi-domain exostosin might mark an evolutionary milestone in this respect. However, although I was able to detect three bi-domain exostosins in the genome of the poriferan sponge *Amphimedon queenslandica* (with homology to EXT1, EXT2, and EXTL3, respectively), HS is not believed to be present in this metazoan lineage, and is thought to have arisen later with the emergence of the Cnidaria (Yamada *et al*, 2011). This either indicates that these enzymes synthesise a completely different polysaccharide, or indicates that they synthesise a GAG or GAG-like polysaccharide that has not yet been successfully detected in these species. Nevertheless, in exhaustive searches, I was not able to find any GT47–GT64 bi-domain enzymes in any earlier-diverging eukaryotes than the Porifera (data not shown). Interestingly, however, I did find that the only GT47-G member in the choanoflagellate *M. brevicollis* (Monbr1|21955) contains a predicted N-terminal sulphatase domain (Pfam: PF00884). Many sulphatases in 'higher' metazoans are involved in the metabolism of sulphated glycosaminoglycans (Hanson *et al*, 2004); hence, it is possible that this *M. brevicollis* enzyme is also involved in modifying a sulphated polysaccharide.

Even in HS-producing animals, the precise activities of these bi-domain enzymes are surprisingly convoluted, with GlcNAcT-I, GlcNAcT-II, and GlcAT-II activities typically distributed amongst orthologues of EXT1, EXT2, and EXTL3 (Busse-Wicher *et al*, 2014). Together, EXT1 and EXT2 are thought to form a bifunctional complex responsible for GlcAT-II and GlcNAcT-II activities (Busse-Wicher *et al*, 2014); however, EXTL3 is also thought be bifunctional in itself, possessing both GlcNAcT-I and GlcNAcT-II activities (Yamada, 2020). To add to this complexity, my results indicated an unambiguous GlcAT-II activity from our preparations of EXTL3ΔN. However, these results are difficult to reconcile with those of previous investigations, in which this activity could not be detected (Kim *et al*, 2001). Moreover, they are at odds with my own structural data, which suggested that the EXTL3 GT47 domain is unable to bind the relevant substrates for such an activity. It also seems unlikely that this inverting activity could originate from the GT64 domain of EXTL3 given that GTs in this family normally use a retaining mechanism—though it has been proposed that the conformation of the GT64 active site is closely related to that of the inverting family GT43 (Taujale *et al*, 2020). Nevertheless, the fact that this activity was not fully inhibited by EDTA—which should prevent nucleotide-sugar binding to the GT64 domain active site—appears to confirm that the observed activity can only originate from a GT47 domain. Hence, the best

explanation for the observed GlcAT-II activity is that it arises from the very low levels of EXT1 and EXT2 protein detected in the proteomic analyses. However, if this explanation is correct, it calls into question the source of the observed GlcNAcT-II activity—as this too might therefore also derive from EXT1 and EXT2. Nevertheless, the fact that the level of GlcNAcT-II was apparently higher than that of GlcAT-II activity suggests that EXTL3 has at least some role in the former, especially given that, at least *in vitro*, the EXT1:EXT2 complex is thought to possess much more GlcAT-II activity than GlcNAcT-II activity (McCormick *et al*, 2000; Busse & Kusche-Gullberg, 2003). Furthermore, the GlcNAcT-II activity of EXTL3 is thought to be weaker than that of the EXT1:EXT2 complex (Yamada, 2020), which may help to explain why the GlcNAcT:GlcAT ratio is not higher.

The most important function of EXTL3, however, appears to lie in its GlcNAcT-I activity—*i.e.* the addition of the first GlcNAc residue to the tetrasaccharide linkage region (itself *O*-linked to the core protein) (Yamada, 2020). While I was unable to assay this activity, or solve a suitable acceptor-bound structure, I attempted to use the apo-EXTL3 structure to explain how the GlcNAcT-I acceptor might be recognised. Previous analysis of the *Mm*EXTL2 structure did not identify any specific interactions between the GT64 domain and the galactosyl residues in the tetrasaccharide linker; hence, only residues that recognise the GlcNAc residue at the acceptor's non-reducing terminus have been well characterised (Pedersen *et al*, 2003). Nevertheless, I speculated that residues within the C-terminal loop of the EXTL3 GT64 domain could possess sidechains with complementary properties to those expected to be enriched at HS attachment sites in the underlying core protein. Because these C-terminal residues are closer to the active site of the opposing monomer than they are to the active site in the same chain, the involvement of these residues in acceptor binding would suggest a role of homodimerisation in the activity of the enzyme.

Less speculatively, I was able to draw some clear conclusions from my structural data regarding the EXTL3 GT47 domain. The EXTL3 structure confirms that GT47 glycosyltransferases (and therefore perhaps GTs from the related GT110 family) adopt a GT-B fold, albeit in a 'streamlined' form that lacks some of the secondary structure elements present in other GT-B families (sequence alignments did not indicate the presence of additional structural elements in other GT47s). I identified a conserved Asp/Glu residue situated in the Cα4 helix that is likely responsible for binding the ribose of the nucleotide sugar in active GT47 enzymes. The importance of this residue is reflected in the fact that expression of IRX10 with a point mutation at this position (E293Q) in otherwise wild-type Arabidopsis plants results in a strong,

dominant-negative *irx* phenotype (Brandon *et al*, 2020). The presence of this structural motif in GT47 (in addition to previously characterised families) may suggest that it is more widespread in the GT-B superfamily than previously thought.

The GT47 domain of EXTL3, in particular, appears to constitute an anomaly within the GT47 family, however. Based on the structural data, it appears that the domain has been inactivated by at least two modifications to the active site: the N-terminal extension of the Cα4 helix and the introduction of an arginine-mediated salt bridge to the Cα4 glutamate. The fact that the GT47 domain did not appear to bind UDP supports the idea that these features block nucleotide sugar binding. This raises the question as to why the GT47 domain is present at all—especially as it appears to contribute little to overall homodimerisation. That said, EXTL3 also possesses a long predicted coiled coil in its stem region that distances the catalytic domain from the membrane (Zak *et al*, 2002). While the purpose of such a feature is not entirely clear, it is possible that the existence of the GT47 domain keeps the GT64 domain at an optimal position for its activity in the Golgi lumen. Furthermore, it has also been proposed that the coiled coil could potentially mediate interactions with other EXTs (Zak *et al*, 2002); hence, it is possible that the GT47 domain could fulfil a similar role in protein–protein interactions.

From a wider perspective, the EXTL3 structure may offer several important pieces of insight into Golgi GTs in general. Firstly, it is clear that this protein forms a symmetric homodimer (a fact that was confirmed prior to the imposition of C2 symmetry during particle reconstruction), with stem regions located suitably for the formation of a coiled coil. The fact that homodimerisation is stabilised by so many features in this enzyme suggests that its oligomeric state could be important for its function. Indeed, many other Golgi GT catalytic domains have been shown to adopt a homodimeric structure (Harrus *et al*, 2018). However, until now, such structures have been solved by crystallography—hence, complete confidence in their oligomeric state has been marred by the potential for crystallographic artefacts (Harrus *et al*, 2018). The cryo-EM structure of EXTL3, which agrees well with the crystallographic structure of *Mm*EXTL2, therefore potentially represents the first homodimeric GT structure to be characterised independently via a non-crystallographic technique.

Furthermore, the EXTL3 structure lends insight into the structure and purpose of bi-domain Golgi GTs. Assuming that the domain organisation can be extrapolated to the structure of other exostosins, it seems that the distance between GT47 and GT64 active sites in these enzymes is too large to permit the direct transfer of substrates between them. Therefore, it appears that

these domains may have been fused to increase efficiency simply by increasing the local concentration of substrate. Whether this is the case for EXT1, EXT2, and other bi-domain Golgi GTs such as the chondroitin synthases and LARGE enzymes will require further experiments to confirm their structures (most likely by cryo-EM). Nevertheless, these results constitute an early step in understanding how glycosyltransferase reactions are organised within the Golgi.

Chapter 5 : Nucleotide sugar specificity of xylan glucuronic acid pyranosyltransferases

5.1 Introduction

Whereas animals possess only a handful of glycosyltransferases from the GT47 family, plant genomes typically encode dozens of GT47-family enzymes, with a large variety in activity (Geshi *et al*, 2011). Particularly striking is the diversity in nucleotide sugar specificity amongst enzymes from GT47 clade A (GT47-A), which encompasses galactosyltransferases, galacturonosyltransferases, arabinofuranosyltransferases, and arabinopyranosyltransferases—perhaps in addition to enzymes with other activities (Pauly & Keegstra, 2016; Zhu *et al*, 2018; Yu *et al*, 2021a, 2021b). The structural characterisation of an animal GT47 domain (see **Chapter 4**) now presents an opportunity to rationalise this diversity from a structural perspective.

In contrast to Arabidopsis secondary cell wall xylan, which exhibits 'unsubstituted' glucuronic acid decorations, the backbone of Arabidopsis *primary* cell wall xylan is decorated with

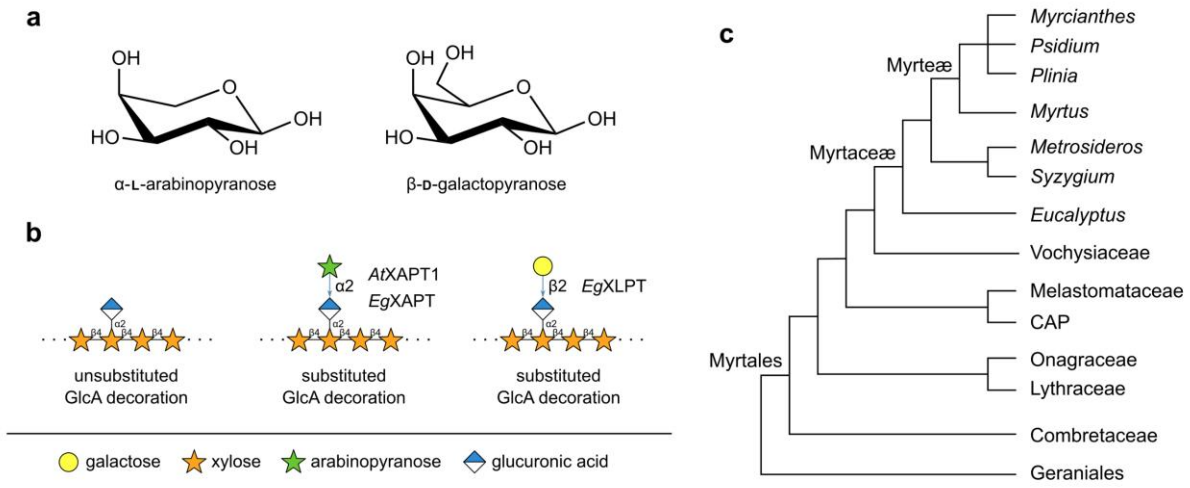


Figure 5.1 Structures of 'substituted-glucuronic-acid' xylan sidechains in eudicots, and the relationship of *Eucalyptus* to other Myrtales plants. a Chemical structures of α-L-arabinopyranose (α-arabinopyranose) and the related monosaccharide β-D-galactopyranose (β-galactose). **b** Substituted and unsubstituted glucuronic acid disaccharide xylan decoration structures and the enzymes responsible for the transfer of arabinopyranose and galactose to glucuronic acid. **c** Simplified cladogram showing the relationships of various taxa within the Myrtales order.

‘substituted-glucuronic-acid’ disaccharide branches, which are thought to possess the structure Ara*p*-α1,2-GlcA-α1,2- (Mortimer *et al*, 2015). Recently, the GT47 enzyme responsible for transferring the α-arabinopyranose (α-Ara*p*) to xylan-linked glucuronosyl residues, xylan arabinopyranosyltransferase 1 (*At*XAPT1), was identified (**Figure 5.1a,b**) (Yu *et al*, 2021b). In the same work, two related enzymes from *Eucalyptus grandis*, *Eg*XAPT and *Eg*XLPT, were identified, and shown to decorate xylan-linked glucuronic acid with α-Ara*p* or β-galactopyranose (β-Gal), respectively. *Eucalyptus grandis* is a member of the Myrtaceæ family, which belongs to the Myrtales order (**Figure 5.1c**); however, the prevalence of such galactosylated xylan branches within these groups has not been studied, and the origin of *Eg*XLPT is unknown. Nevertheless, with their high similarity in sequence and function, this pair of enzymes therefore represents a potentially useful model for understanding how subtle changes in glycosyltransferase structure can bring about changes in nucleotide sugar specificity.

However, the prediction and rationalisation of GT donor substrate specificity is not always straightforward—the determinants of nucleotide sugar recognition often appear to be subtle (sometimes distal to the binding pocket) and without any general trends (Lairson *et al*, 2008; Chang *et al*, 2011). Nevertheless, some progress has been made in the past in understanding the donor specificity of particular enzymes and GT families. An early example is constituted by the work on the ABO(H) blood-group-determining enzymes ‘GTA’ and ‘GTB’ (not to be confused with the names of the GT-A and GT-B folds), which catalyse the transfer of a β-GalNAc or β-Gal residue to the H antigen, respectively. These two GT-A-fold enzymes (which derive from two different *ABO* alleles) were found to differ by only four amino acids (Yamamoto *et al*, 1990)—expediting experiments into their specificities. Subsequently, by making point mutation combinations, it was determined that two of these amino acid changes (Leu/Met266 and Gly/Ala268) are both necessary and sufficient for a full switch between β-*N*-acetylgalactosaminyltransferase activity and β-galactosyltransferase activity (Yamamoto & Hakomori, 1990; Seto *et al*, 1999; Kamath *et al*, 1999). Subsequently solved crystal structures revealed that these two residues directly contact the donor sugar, with the sidechain of Leu/Met266 in particular helping to distinguish between the acetamido and hydroxyl groups at C2 of the donor sugar, respectively (Patenaude *et al*, 2002).

More recently—and more relevantly to XAPT/XLPT structure-function, progress has been made in understanding the donor specificities of GT1 family members, which adopt a GT-B fold. These enzymes exhibit an enormous variety of acceptor and donor specificities

(particularly so in plants (Jones & Vogt, 2001)), and are often individually promiscuous (Biswas & Thattai, 2020). Nevertheless, donor sugar recognition is thought to be mediated by residues at the N-terminus of Cα5 (see **Figure 1.7** for GT-B secondary structure nomenclature), as well as potentially any of the Nβ1–Nα1 loop, the Nβ5–Nα5 loop, the Cβ1–Cα1 loop, or the interdomain linker (Osmani *et al*, 2009). In somewhat of a breakthrough in computational prediction techniques, the acceptor and donor specificities of various uncharacterised GT1s have been successfully predicted by using a decision tree algorithm trained on a comprehensive library of characterised Arabidopsis GT1 activities (Yang *et al*, 2018). Even more recently, machine learning techniques have been applied to the entire GT-A family, with promising results (Taujale *et al*, 2020). However, both experiments made use of the large number of characterised activities in the GT1 and GT-A families, respectively. Although a good deal of activities have been identified in the GT47 family (see *Section 1.4.2*), they may not yet be of sufficient number for such approaches.

Nevertheless, I attempted to investigate the difference in activity between *Eg*XAPT and *Eg*XLPT by analysing their protein sequences, modelling the XAPT protein structure, studying the prevalence of β-galactosylated xylan amongst plants, and using site-directed mutagenesis to probe the relevance of individual amino acids. In this chapter, I reveal the detailed relationships between enzymes in the GT47-A clade, especially those related to XAPT, through a series of phylogenies and sequence alignments. I also show that the Ala235→Gly change between *Eg*XAPT and *Eg*XLPT is neither necessary nor sufficient to bring about a change in nucleotide sugar specificity, contrary to an initial structure-based hypothesis. Furthermore, I show that galactosylated xylan sidechains are widespread in the Myrtaceæ family (to which *E. grandis* belongs), and that XAPT and XLPT likely diverged from one other in a recent ancestor of this family.

5.2 Results

5.2.1 Clade A of CAZy family GT47 comprises at least seven subgroups

GT47 clade A contains several different enzymes in Arabidopsis, including not only XAPT1, but also MUR3, XLT2, XUT1, MBGT1, and six other uncharacterised proteins (GT11, GT12, GT13, GT15, GT19, and GT20) (Li *et al*, 2004). To determine how the functions of these enzymes might have arisen through evolution, and to identify further potential orthologues of *At*XAPT1, I wanted to construct a comprehensive large-scale GT47-A phylogeny for the entire plant kingdom. Therefore, I downloaded MUR3-related protein sequences from the genomes, transcriptomes, and proteome models of eudicot, monocot, gymnosperm, and other plant

species using the plant comparative genomics platform PLAZA. Using HMMER (Eddy, 2011), a GT47-A hidden Markov model (HMM) was constructed and used to search for additional GT47-A sequences in several further phylogenetically useful species, including two fern species and the streptophyte alga *Klebsormidium nitens*. In total, 1,163 sequences were collected from 97 species, including representatives of every major phylogenetic group within the embryophytes (see **Table 2.1** for a full list of species). Using the EXTL3 structure as a guide, these sequences were truncated to their GT47 domain (corresponding to residues 196–538 in EXTL3). After alignment, a phylogeny was constructed using FastTree (Price *et al*, 2010), with 100 bootstrap replicates. The resultant tree revealed the existence of seven main subgroups within GT47-A: a subclade containing the two *Physcomitrium patens* enzymes *Pp*XLT2 and *Pp*XDT but no 'higher plant' sequences ('group I'), an XUT1/GT20-related subclade ('group II'), an XLT2-related subclade (III), a GT19-related subclade (IV), a XAPT1-related subclade (V), a MUR3-related subclade (VI), and an MBGT1/GT12/GT13/GT15-related subclade (VII) (**Figure 5.2**). Sequences from *Klebsormidium nitens* were monophyletic, and this small subclade was tentatively assigned as the root of the tree. Interestingly, a large group of GT47-A enzymes specifically from the Poaceæ family (grasses) was also observed to form a distinct subclade, apparently diverging from the ancestor of group VI (the MUR3-related subclade). However, the overall phylogeny was not well resolved at the bifurcation of groups VI and VII, and, owing to the lack of bootstrap support and the implausibility of the implied ancestry, the exact placing of this Poaceæ-specific subclade seems doubtful in this case. Nevertheless, this subclade included the recently characterised *Sb*GT47_2 enzyme from *Sorghum bicolor*, which was shown to act as a galactosyltransferase to xyloglucan (Xu *et al*, 2018); hence, it is possible that this subclade may have emerged from group VI itself. The previously characterised xylan-acting enzymes *At*XAPT1, *Eg*XAPT, and *Eg*XLPT, on the other hand, were unambiguously grouped together in group V. Interestingly, Poaceæ family plants did not appear to exhibit any sequences in this group; nor did many monocots exhibit sequences in group II. As a whole, these results represent the first comprehensive pedigree of the GT47-A clade and facilitate the identification of GT47-A orthologues from a wide range of plant species.

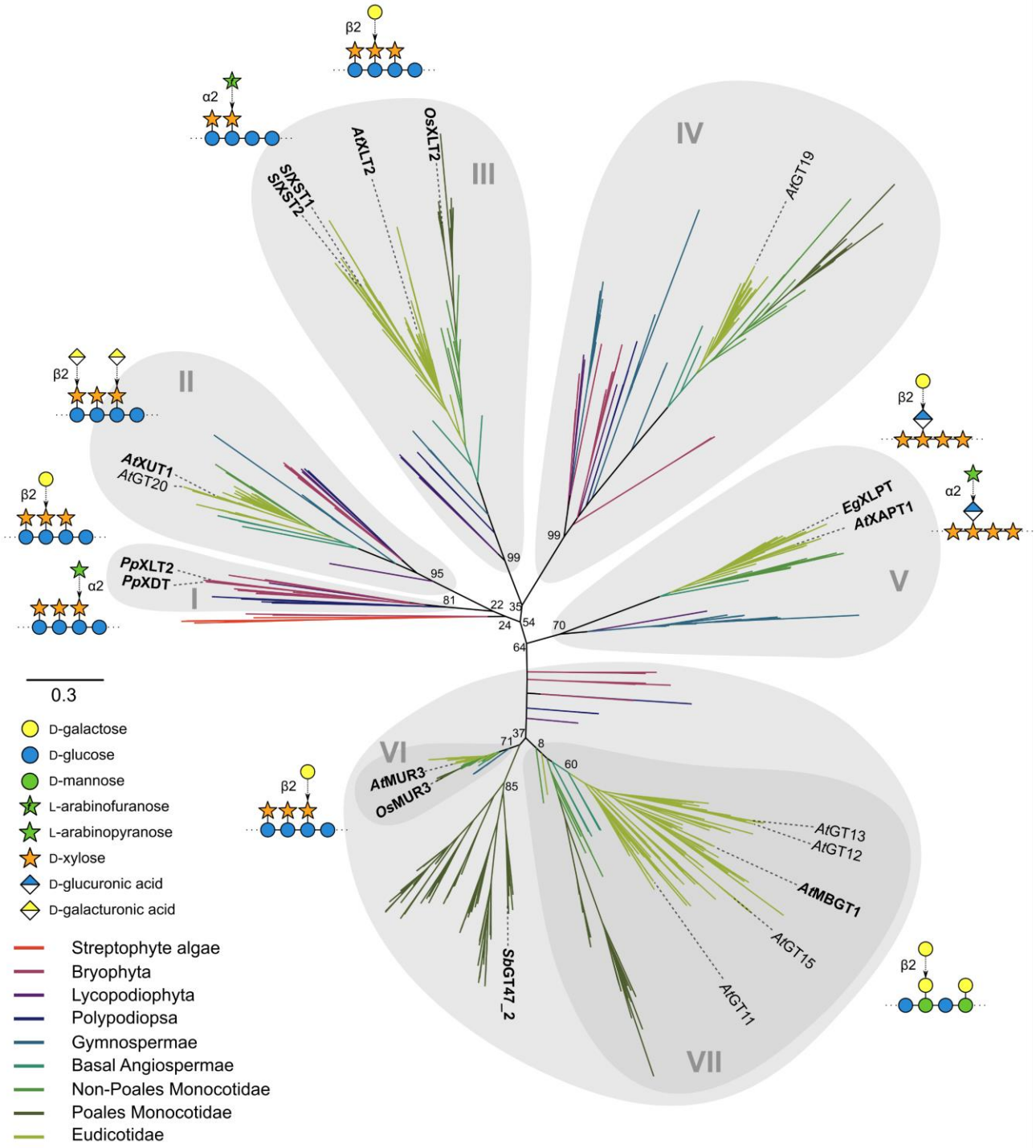


Figure 5.2 Phylogeny of glycosyltransferase family GT47, subclade A. FastTree was used to construct a phylogeny from 1,163 GT47-A protein sequences from a variety of plant species. The streptophyte algæ and Lycopodiophyta are represented by only one species each: *Klebsormidium nitens*, and *Selaginella moellendorffii*, respectively. Arabidopsis GT47-A members, as well several other characterised GT47-A enzymes, are labelled on the tree. Bootstrap values are displayed only for major branchpoints.

5.2.2 Both XAPT and XLPT genes are present in Myrtaceæ family genomes but XLPT appears to be absent in the wider Myrtales

Xylan decorated with galactosylated glucuronic acid disaccharides, or 'galactoglucuronoxylan', was first identified in *Eucalyptus* spp.; however, the prevalence of this decoration amongst eudicots has not been fully investigated (Shatalov *et al*, 1999; Peña *et al*, 2016). In *Eucalyptus grandis*, the synthesis of these decorations likely involves the galactosyltransferase *Eg*XLPT, a close relation of *Eg*XAPT (Yu *et al*, 2021b). Therefore, the ability of plants to make galactoglucuronoxylan is likely to require the possession of a XLPT orthologue. To estimate when XLPT-related enzymes might have first arisen in eudicot evolution, I used the group V sequences extracted from the GT47-A tree (see above) to build a new HMM focused specifically on XAPT-family sequences. This HMM was used to identify XAPT homologues in published proteome models for species in the Myrtaceæ family (to which *E. grandis* belongs) and the wider Myrtales order. For the Myrtaceæ family, *Metrosideros polymorpha*, *Syzygium oleosum*, and *Rhodamnia argantea* proteome models were used. *Punica granatum* (pomegranate) offered the only available proteome model from the rest of the Myrtales order; for this reason, I also used TBLASTN to search for XAPT homologues in Myrtales transcriptomes available through the 1,000 Plants initiative (Leebens-Mack *et al*, 2019); subsequently, I included all close hits from *Oenothera rosea*. With the supplementary Myrtales sequences included, the group V protein sequences (106 in total) were truncated to their GT47 domain and aligned. *At*MUR3 was also included for rooting purposes. A phylogeny was then constructed using RAxML. The tree (trimmed to relevant species for ease of viewing) showed that most eudicot species (including Arabidopsis) possess only one XAPT homologue in their genome (**Figure 5.3**). Conspicuously, however, homologues in the Myrtaceæ family comprised two subgroups, with *Eg*XAPT placed in one subgroup and *Eg*XLPT placed in the other. In contrast, *P. granatum* and *O. rosea* appeared to possess only one XAPT homologue each, both of which were placed in the *Eg*XAPT subclade. If this topology is correct, the data imply that a XAPT homologue was duplicated to form the predecessors of *Eg*XAPT and *Eg*XLPT in an ancestor or early progenitor of the Myrtales, but that the XLPT-type gene was subsequently lost from the Lythraceæ/Onagraceæ group to which *P. granatum* and *O. rosea* both belong. More conservatively, these data indicate that the XAPT- and XLPT-related groups must have diverged at least before the emergence of the Myrtaceæ as a distinct family.

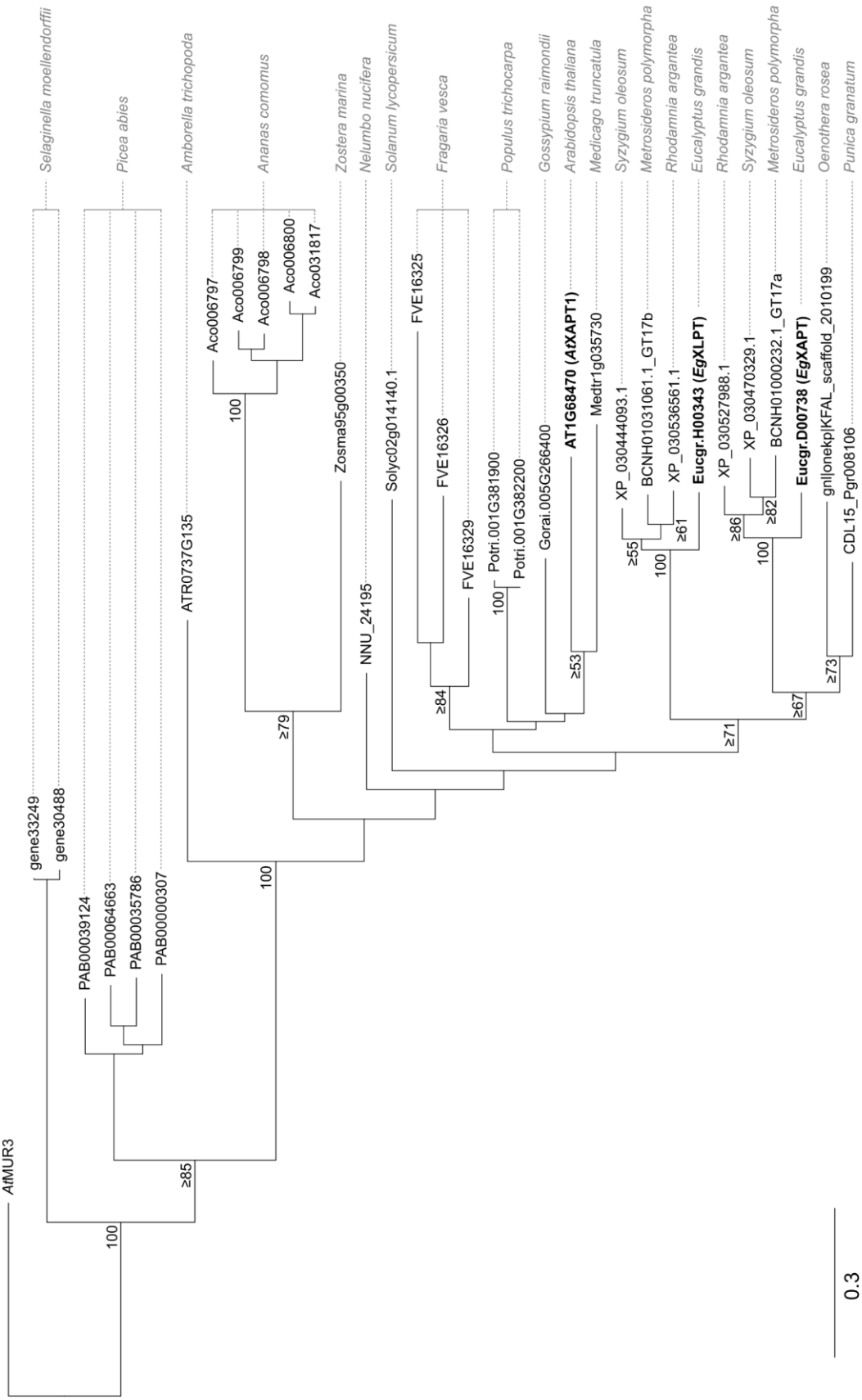


Figure 5.3 Trimmed phylogeny of GT47-A group V (including related sequences from Myrtales plants). One hundred and six XAPT-subgroup sequences, either extracted from the tree shown in **Figure 5.2** or obtained from Myrtales genomes using a XAPT HMM, were aligned and a phylogeny was constructed using RAxML. Branches were then removed to leave only a subset of species for clarity; hence, the bootstrap values from the original tree are underestimates of the splits shown in this reduced tree. Only bootstrap values above 50 are shown.

5.2.3 Eg*XAPT and* Eg*XLPT exhibit only a small number of potential structural differences close to the predicted donor sugar binding site*

To gain insight into how XAPT-related glycosyltransferases achieve nucleotide sugar specificity, I constructed a structural model of the *At*XAPT1 GT47 domain using the I-TASSER server, submitting the GT47 domain of EXTL3 as a threading template. The best-scoring model exhibited the expected double Rossmann-fold structure, with one fewer turn to the Cα4 helix compared with EXTL3 (as predicted; see *Section 4.2.6*). Next, I aligned the *At*XAPT1 model with the EXTL3 model, which had itself been aligned to other similar GT-Bs using the DALI server (see *Section 4.2.6*). This revealed the location of various nucleotides/nucleotide sugars bound to said GT-Bs relative to the aligned *At*XAPT1 model (see **Figure 5.4a–c**). These crudely positioned substrates occupied a similar position as they did for EXTL3, except that the donor sugar portion no longer clashed with the proximal end of the Cα4 helix (as predicted—see *Section 4.2.6*) and was surrounded by the N-terminus of Cα4, the Nβ5–Nα5 loop, and the proximal portion of the Nβ7–Cα1 loop. In addition, sidechains at the N-terminus of Cα1 appeared to be in close proximity to the phosphates of the nucleotide sugar; however, this prediction seemed spurious as the N-terminus of Cα1 in EXTL3 is roughly 20 Å away from the position of the nucleotide phosphates and is separated from this site by several other secondary structure elements. Conversely, although the Nβ6–Nβ7 and Cβ2–Cα3 loops come close to the nucleotide sugar binding pocket in EXTL3, these loops were not predicted to do so in the *At*XAPT1 model, though the strength of this prediction is unclear. Hence, I paid close attention to the Nβ5–Nα5 loop, the Nβ6–Nβ7 loop, the Nβ7–Cα1 loop, the N-terminus of Cα1, and the N-terminus of Cα4 in the following analyses.

Given the relation in sequence and activity between *Eg*XAPT and *Eg*XLPT, I reasoned that it might be possible to correlate their difference in activity with a particular difference in protein sequence. To that end, I aligned the two proteins using MUSCLE. The two sequences showed relatively high similarity, with 50 % shared identity (**Figure 5.4d**). I attempted to construct a model for *Eg*XAPT using I-TASSER, but unfortunately, for unknown reasons, the output model was untenable. Therefore, instead, I aligned the protein sequences of *Eg*XAPT and *Eg*XLPT to that of *At*XAPT1, and mapped amino acid differences between the first two proteins onto the model of the third (**Figure 5.4a**) Remarkably, only two sequence differences between *Eg*XAPT and *Eg*XLPT were apparent in the vicinity of the putative donor sugar binding site. The closest of these constituted a single amino acid substitution—Ala235 *vs*

Gly212 (*Eg*XAPT *vs Eg*XLPT, respectively) in the Nβ5–Nα5 loop—while the other, situated near the N-terminus of Cα1, constituted a single residue insertion and altered nearby amino acid identities in *Eg*XLPT in comparison to *Eg*XAPT (exact residues are not given because alignments were not consistent in this region). Since I judged the prediction that the Cα1 N-terminus comes close to the active site to be spurious, I considered the Ala→Gly change to be the more likely explanation for the difference in activity. However, differences in amino acid identities were also observed in the Cβ2–Cα3 loop; hence, if (in contradiction with the *At*XAPT1 model) this third structural component *is* in fact in the vicinity of the donor sugar, as one might predict from the EXTL3 structure, these residues could also constitute candidates for such a role in substrate specificity. In contrast, no amino acid differences were observed in the Nβ6–Nβ7 loop or around the N-terminus of Cα4.

5.2.4 Mutation of Ala235 in Eg*XAPT to glycine is insufficient to alter enzyme substrate specificity*

The idea that an Ala→Gly change could bring about a change in binding of L-Ara*p*→D-Gal is an attractive hypothesis, because the former constitutes a loss of a methyl group that might compensate sterically for the latter's gain of a methoxy group (L-Ara*p* and D-Gal being structurally analogous—see **Figure 5.1**). To investigate the Ala235 *vs* Gly212 difference further, I inspected the sequence alignment used for the XAPT phylogeny above to assay the conservation of amino acid identity at this position in the Nβ5–Nα5 loop (in addition to the conservation of amino acids near the Cα1 N-terminus and amino acids in the Cβ2–Cα3 loop). Interestingly, barring a number of sequences from the Rosaceæ family, Ala/Ser is strongly conserved at this position in XAPT-related enzymes, thus making the apparently unique change to Gly in *Eg*XLPT somewhat conspicuous (**Figure 5.5**). In contrast, the residues in the Cβ2–Cα3 loop were not at all conserved amongst XAPT homologues, suggesting that these residues are not involved in nucleotide sugar binding. Since I had also rejected the idea that the Cα1 N-terminus could be involved in nucleotide sugar binding, I hypothesised that the replacement of Ala235 in *Eg*XAPT with glycine in *Eg*XLPT could be responsible for the difference in activity.

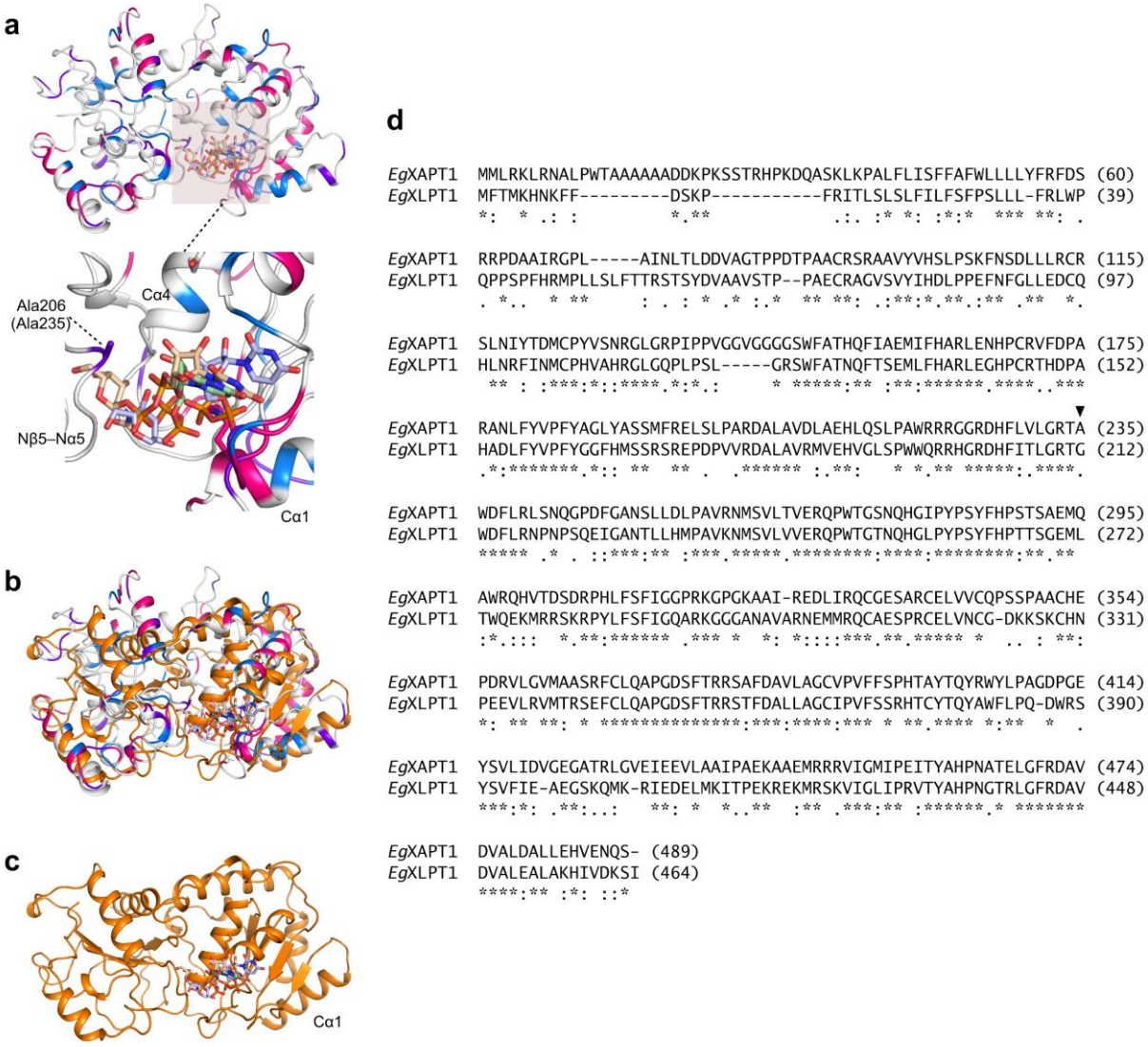


Figure 5.4 Structural model of *At*XAPT1 and sequence differences between *Eg*XAPT and *Eg*XLPT. a I-TASSER model of *At*XAPT1, using the EXTL3 GT47 domain as a template. Differences between *Eg*XAPT and *Eg*XLPT, mapped onto this structure, are shown by the following colours: white (identical), blue (minor difference), purple (intermediate difference), hot pink (major difference). Donor substrates shown are UDP and Glc from *Dm*PoGLUT1 (PDB: 5F84; pale blue), UDP from T4 phage BGT (PDB: 1IXY; pale green), and UDP-GalNAc from *Cj*PglH (PDB: 6EJI; buff). The position of Ala206, which corresponds to Ala235 in *Eg*XAPT, is indicated. **b** Same as **a** but with EXTL3 GT47 domain aligned. **c** EXTL3 and donor substrates only. **d** MUSCLE protein sequence alignment of EgXAPT and EgXLPT. Ala235 and Gly212 are indicated in the alignment by a black arrowhead.

	Nβ5–Nα5 loop	Nβ7–Cα1 loop / Cα1 N-terminus	Cβ2–Cα3 loop
AT2G20370 (AtMUR3)	RITWDFRR	R-PDN----------PK	FGES--KCHA
gene33249	RITWDFIR	K-A----------KN	-GESGGECHD
gene30488	RITWDFIR	K-A----------KN	-GESGGECHD
PAB00039124	RIAWDFVR	R-PGLDKAFSFAGAPRPGLDKA	QGPS--ICHE
PAB00064663	RIAWDFVR	R-PGLD----------KA	QGPS--ICHD
PAB00035786	RIAWDFVR	R-PGLD----------KA	QGPS--ICHN
PAB00000307	RIARDFMR	R-PRLE----------RA	KGLS--ICNE
ATR0737G135	RTSWDFMR	RPPGLE----------KA	PGAS--GCYE
Aco006797	RTSWDLMR	R-NGGQ----------KE	-QAS--GCYT
Aco006799	RTSGELMR	R-TKGQ----------KE	RGSK--DCRT
Aco006798	RTSGELMR	RTKGQ----------KE	RGSK--DCRT
Aco006800	RTSWELMR	R-TGGQ----------QE	RGSK--DCRT
Aco031817	RTSWELMR	R-TGGQ----------QE	RGSK--DCRT
Zosma95g00350	RTAWDFMR	R-KGLG----------KA	PGSV--LCYE
NNU_24195	RTAWDFMR	R-TGVE----------KA	VGAS--KCHE
Solyc02g014140.1	RTAWDFMR	R-TGKK----------KV	HGPS--KCHN
FVE16325	--------	----------	----------
FVE16326	RVVRDFKS	R-R----------KA	DEAR--KCQE
FVE16329	RTAWDFMR	R-KGLE----------KA	NGAS--KCHE
Potri.001G381900	RTAWDFLR	R-RGVE----------KA	KGPS--KCHY
Potri.001G382200	RTAWDFLR	R-RGVE----------KA	KGPS--KCHD
Gorai.005G266400	RTSWDFMR	R-KGVG----------KA	HGNP--KCYN
AT1G68470 (AtXAPT1)	RTAWDFMR	R-KGLE----------KA	NGGS--RCHN
Medtr1g035730	RTAWDFMR	R-KGLG----------KA	GGNS--KCHQ
XP_030444093.1	RTAWDFLR	RTGGGA----------NA	ANGS--KCHN
BCNH01031061.1_GT17b	RTAWDFLR	R-TGGG----------ANA	ANAS--KCHH
XP_030536561.1	RTAWDFLR	R-TGEG----------ANA	ANAS--KCRH
Eucgr.H00343 (EgXLPT)	RT**G**WDFLR	R**KGGG**A----------**N**A	**DKKS**--**K**CH**N**
XP_030527988.1	RTAWDFLR	R-KGSG----------MA	ASPA--ACHK
XP_030470329.1	RTAWDFLR	R-RGPG----------MA	ASPA--ACHD
BCNH01000232.1_GT17a	RTAWDFLR	R-KGPG----------KA	ASPA--ACHD
Eucgr.D00738 (EgXAPT)	RT**A**WDFLR	R-**K**G**P**G----------**K**A	**SSPA**--**A**CH**E**
KFAL_scaffold_2010199	RTSWDFMR	R-NGPG----------KA	GTGS--KCHE
OWM66937.1	RTAWDFMR	R-KGLE----------KA	GGDS--KCHS

whole alignment:

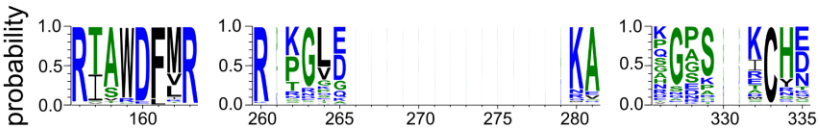


Figure 5.5 Structural elements in XAPT/XLPT that contain sequence differences between *Eg*XAPT and *Eg*XLPT and that are also potentially proximal to the donor sugar binding site. Relevant excerpts from the alignment used to create the tree shown are shown. Only a subset of species is displayed for brevity. Residues differing between *Eg*XAPT and *Eg*XLPT are shown in bold. Also shown are sequence logos presenting amino acid identity probabilities at each position in the overall alignment (created using WebLogo 3).

To test this hypothesis, I generated four independent transgenic Arabidopsis lines expressing an *Eg*XAPT[A235G] point mutant under the control of the IRX3 promoter. This promoter is specific to secondary cell wall-synthesising cells, where Ara*p*/Gal-substituted glucuronic acid is normally totally absent from xylan. In order to characterise the resultant xylan structures in these plants, I prepared AIR material from the bottom portion of their stems. I treated these samples, as well as AIR from plants expressing wild-type *Eg*XAPT or *Eg*XLPT (provided by Dr Li Yu), with a (substituted-GlcA-compatible) GH30 xylanase, followed by GH115 α-glucuronidase to remove unsubstituted glucuronic acid. GH30 xylanase obligately binds glucuronic acid decorations at the −2 position relative to cleavage, thus generating a ladder of acidic oligosaccharides whose length reveals the spacing between decorations (Malgas *et al*, 2019; Bromley *et al*, 2013). PACE analysis indicated that, as expected, digestion of the xylan from all transgenic lines produced not only simple xylo-oligosaccharides, but also—in contrast to the wild-type control (WT)—xylo-oligosaccharides decorated with substituted glucuronic acid (identified with reference to the migration reported by Yu *et al* (2021b)), confirming that *Eg*XAPT[A235G] is expressed and active in all four lines (**Figure 5.6**).

Next, in order to distinguish Ara*p* substitutions from Gal substitutions, I tested the sensitivity of these disaccharide decorations to β-galactosidase. A subset of the same AIR samples were digested simultaneously with GH115 α-glucuronidase and GH11 xylanase, which requires unsubstituted xylosyl residues at the +1, −1, and −2 subsites (and additionally cannot tolerate a disaccharide sidechain at the −3 subsite (Mortimer *et al*, 2015)). These enzymes were then removed before addition of GH35 β-galactosidase; the products were analysed by PACE. Based on the previously documented migration of GH11 xylanase products of Arabidopsis primary cell material (Mortimer *et al*, 2015), the products in this experiment were judged to constitute xylose (X), xylobiose (XX), and xylopentaose decorated with either the Ara*p*-α1,2-GlcA-α1,2- or Gal-β1,2-GlcA-α1,2- disaccharide (X(A)UXXX or X(L)UXXX) (**Figure 5.7**). Furthermore, for the *Eg*XLPT-expressing line, treatment with β-galactosidase resulted in a downwards shift of the X(A)UXXX/X(L)UXXX band consistent with the loss of one monosaccharide, indicating that the original GH11/GH115 product contained a terminal galactosyl residue. In contrast, for the wild-type *Eg*XLPT- and *Eg*XLPT[A235G]-expressing lines, the X(A)UXXX/X(L)UXXX band did not show substantial sensitivity to β-galactosidase, indicating that these products do *not* contain significant amounts of terminal galactose. These results clearly demonstrate that mutation of Ala235 to glycine is not sufficient to alter *Eg*XAPT activity.

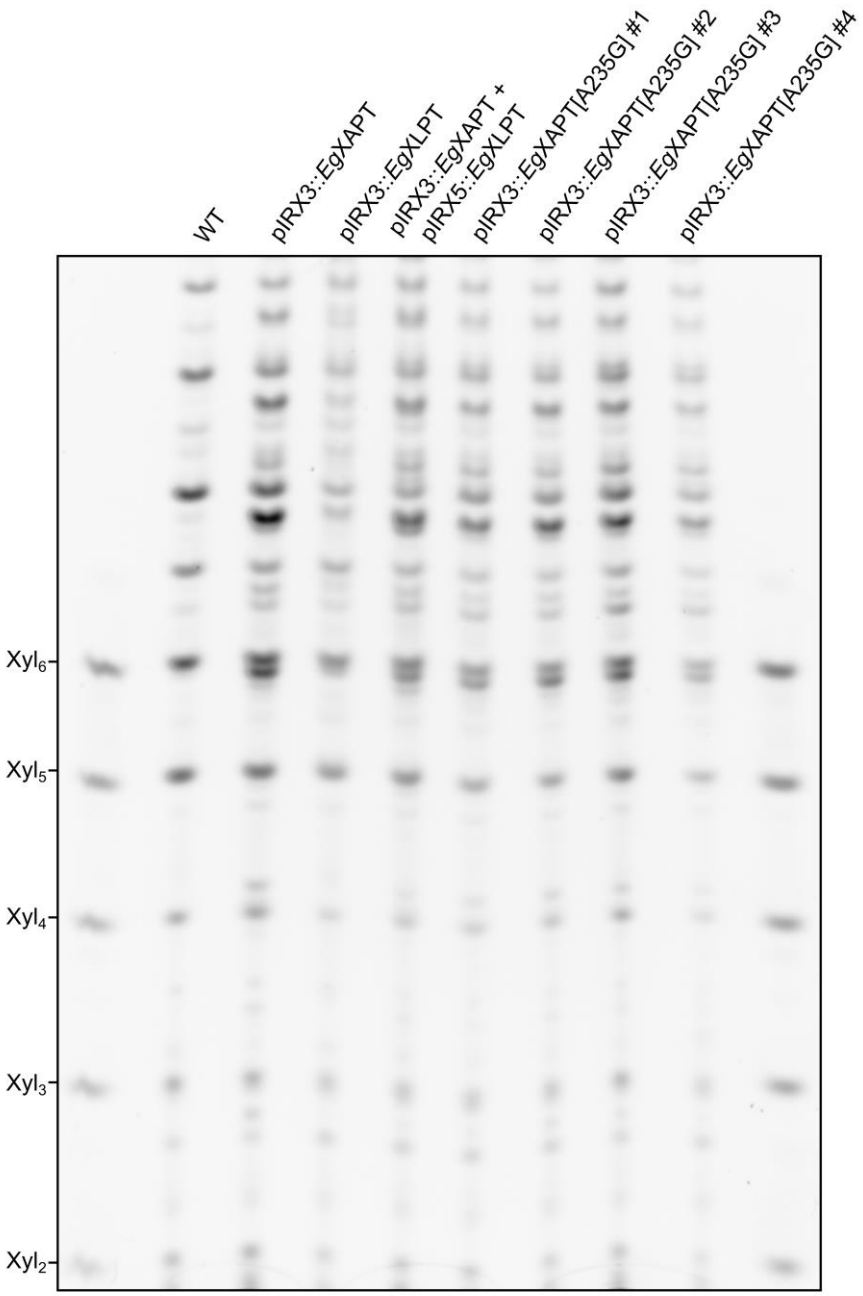


Figure 5.6 Xylan digestion with GH30 *endo*-xylanase and GH115 α-glucuronidase reveals that expression of *Eg*XAPT[A235G] in Arabidopsis bottom stem results in new decorations similar to those introduced by *Eg*XAPT and *Eg*XLPT. Alkali-extracted hemicellulose from bottom-stem AIR of Arabidopsis plants (wild type (WT) and T_1 transgenic lines expressing *Eg*XAPT and/or *Eg*XLPT, or *Eg*XAPT[A235G], all under the IRX3 promoter), was digested with GH30 xylanase from *Dickeya chrysanthemi* P860219, which requires the xylose at the −2 subsite to be decorated with either substituted or unsubstituted glucuronic acid. Products were then treated with α-glucuronidase to remove unsubstituted GlcA residues before ANTS derivatisation and analysis by PACE.

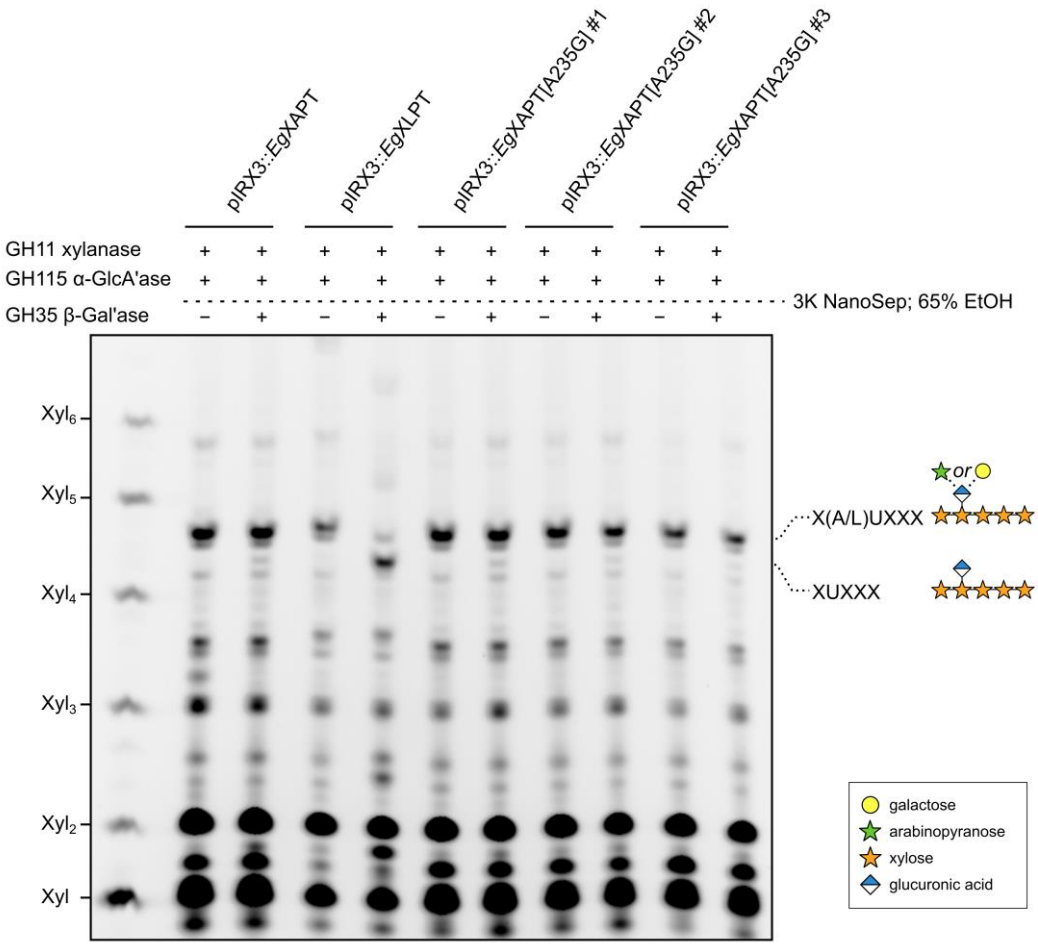


Figure 5.7 Substituted-glucuronic-acid disaccharide xylan decorations created by *Eg*XLPT, but not *Eg*XAPT or *Eg*XAPT[A235G], are sensitive to β-galactosidase. Alkali-extracted hemicellulose from bottom-stem AIR of Arabidopsis plants (WT and T_1 transgenic lines expressing *Eg*XAPT, *Eg*XLPT or *Eg*XAPT[A235G], all under the IRX3 promoter), was digested simultaneously with GH11 *endo*-xylanase and GH115 α-glucuronidase. The products were subsequently treated with GH35 β-galactosidase before ANTS derivatisation and analysis by PACE.

5.2.5 Galactoglucuronoxylan is detectable in many, but not all members of the Myrtaceæ family

Although I found that many Myrtaceæ species possess homologues of XLPT, the possibility still remained that *Eg*XLPT could be the only member of the group to possess galactosyltransferase activity. Unlike in *Eucalyptus* spp., XLPT homologues from the Myrtaceæ-family plants *M. polymorpha*, *S. oleosum* and *R. argantea* were not found to possess the Ala→Gly change seen in *Eg*XLPT; hence, if this amino acid substitution is at least *required* for galactosyltransferase activity, then galactoglucuronoxylan should similarly be unique to

Eucalyptus spp.. Therefore, I attempted to detect galactoglucuronoxylan in other Myrtaceæ species outside of the *Eucalyptus* genus. To this end, I collected young side branches from five additional plants: *Metrosideros excelsa*, *Myrtus communis* (myrtle), *Plinia cauliflora*, *Psidium guajava* (guava), and *Myrcianthes pungens*. Because cell wall composition varies between different wood tissues (Gorshkova *et al*, 2010), the branches were sectioned into periderm/phloem and xylem/pith, and AIR was prepared separately from each. AIR from *E. dalrympleana* tissues and *Eg*XAPT/XLPT-expressing Arabidopsis plants was also provided by Dr Li Yu. To investigate the patterning of unsubstituted and substituted glucuronic acid xylan decorations from these samples, I extracted xylan from each using alkali, and treated with (substituted-GlcA-compatible) GH30 xylanase. The products of GH30 digestion of the Arabidopsis bottom stem and Myrtaceæ xylem samples revealed a wide range of unsubstituted glucuronic acid spacings, with a predominance of evenly spaced decorations (**Figure 5.8**), as previously described for Arabidopsis secondary cell wall xylan (Bromley *et al*, 2013). In addition, the Myrtaceæ samples also exhibited some extra bands with respect to the Arabidopsis wild type xylan that co-migrated with the *Eg*XAPT-overexpressing Arabidopsis line, indicating the presence of *substituted* glucuronic acid in these xylans. However, in contrast to the xylem-tissue xylan, the *E. dalrympleana* and *M. excelsa* phloem/periderm xylan exhibited a structure dominated by a six-residue glucuronic acid spacing typical of primary cell wall xylan in Arabidopsis (Mortimer *et al*, 2015) (the glucuronic acid on this xylan was also substituted to some extent; **Figure 5.9**). Xylan from the other Myrtaceæ phloem/periderm samples, on the other hand, resembled that from the xylem samples, though this could have well been due to the difficulty encountered in sectioning these specimens, which had, in contrast to *E. dalrympleana* and *M. excelsa*, only a thin layer of phloem/periderm.

Next, I attempted to quantify and characterise the substituted glucuronic acid decorations in these xylans. I digested the same periderm/phloem and xylem/pith samples with GH11 xylanase and GH115 α-glucuronidase, followed by GH35 β-galactosidase, before analysing the products by PACE. As above, the products were judged to constitute xylose (X), xylobiose (XX), and, in all but wild-type Arabidopsis, X(A)UXXX or X(L)UXXX (**Figure 5.10–Figure 5.12**). Of the non-transgenic samples, *M. communis* exhibited the highest amounts of substituted glucuronic acid in both periderm/phloem and xylem/pith. For most of the Myrtaceæ samples, as well as the Arabidopsis lines expressing *Eg*XLPT, the X(A/L)UXXX band was sensitive to β-galactosidase, indicating the presence of galactosylated sidechains on these xylans. In contrast, xylanase products from *M. excelsa*, as well as those from Arabidopsis

expressing only *Eg*XAPT, were completely resistant to β-galactosidase treatment, indicating the absence of galactosylated xylan sidechains in these samples. These results demonstrate unambiguously that xylan sidechains are galactosylated in several other Myrtaceæ species besides *Eucalyptus* spp., suggesting that these plants harbour an enzyme with similar activity to *Eg*XLPT. Furthermore, these results suggest that the Ala→Gly change is not necessary for galactosyltransferase activity in XLPT enzymes.

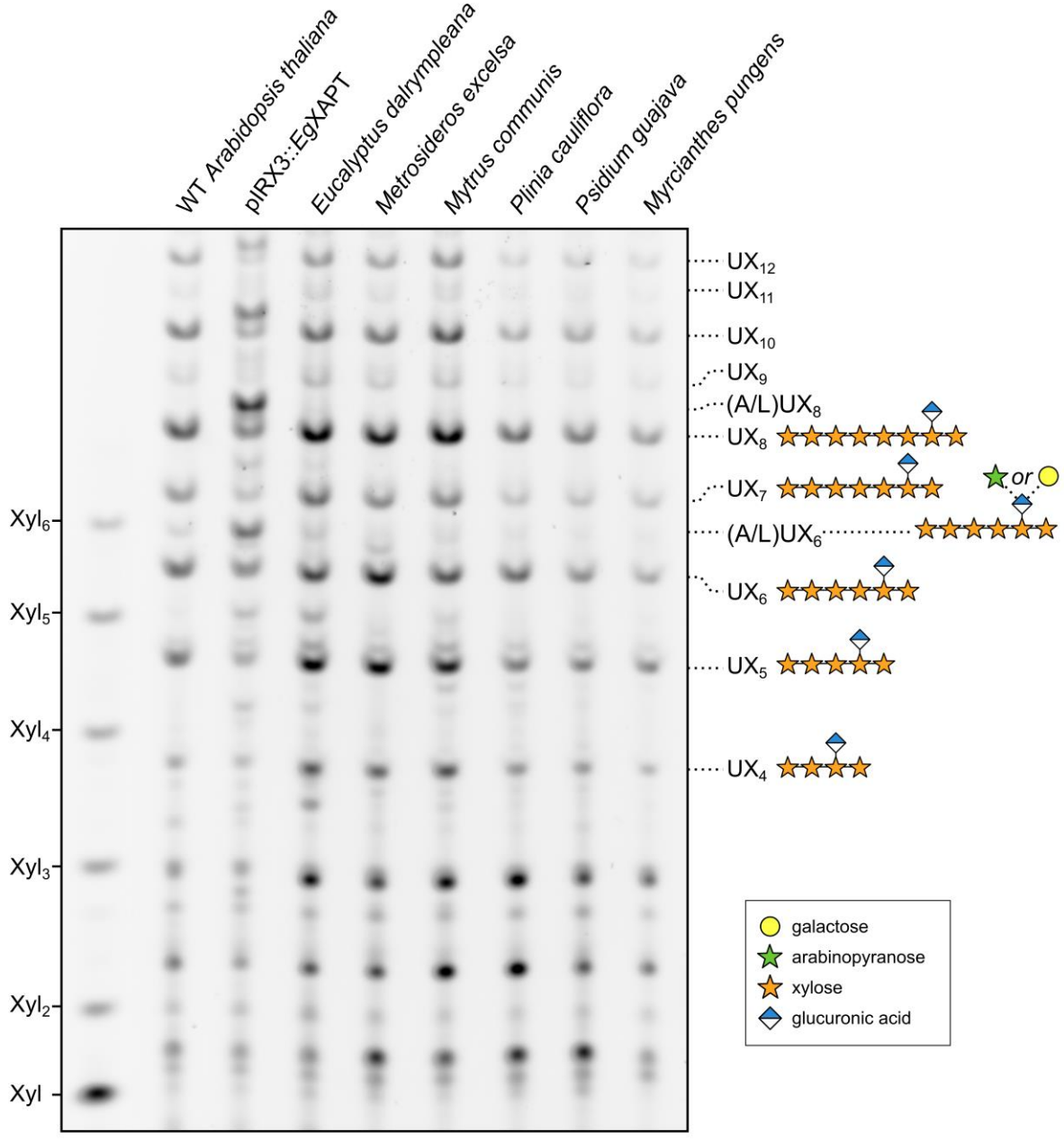


Figure 5.8 Xylan from xylem tissues of Myrtaceæ plants exhibits a preferential even-spacing of glucuronic acid decorations. Alkali-extracted hemicellulose from bottom stem AIR (in the case of Arabidopsis samples) or xylem/pith AIR (in the case of Myrtaceæ samples) was digested with GH30 *endo*-xylanase from *Dickeya chrysanthemi* P860219. The products were derivatised with ANTS and separated by PACE.

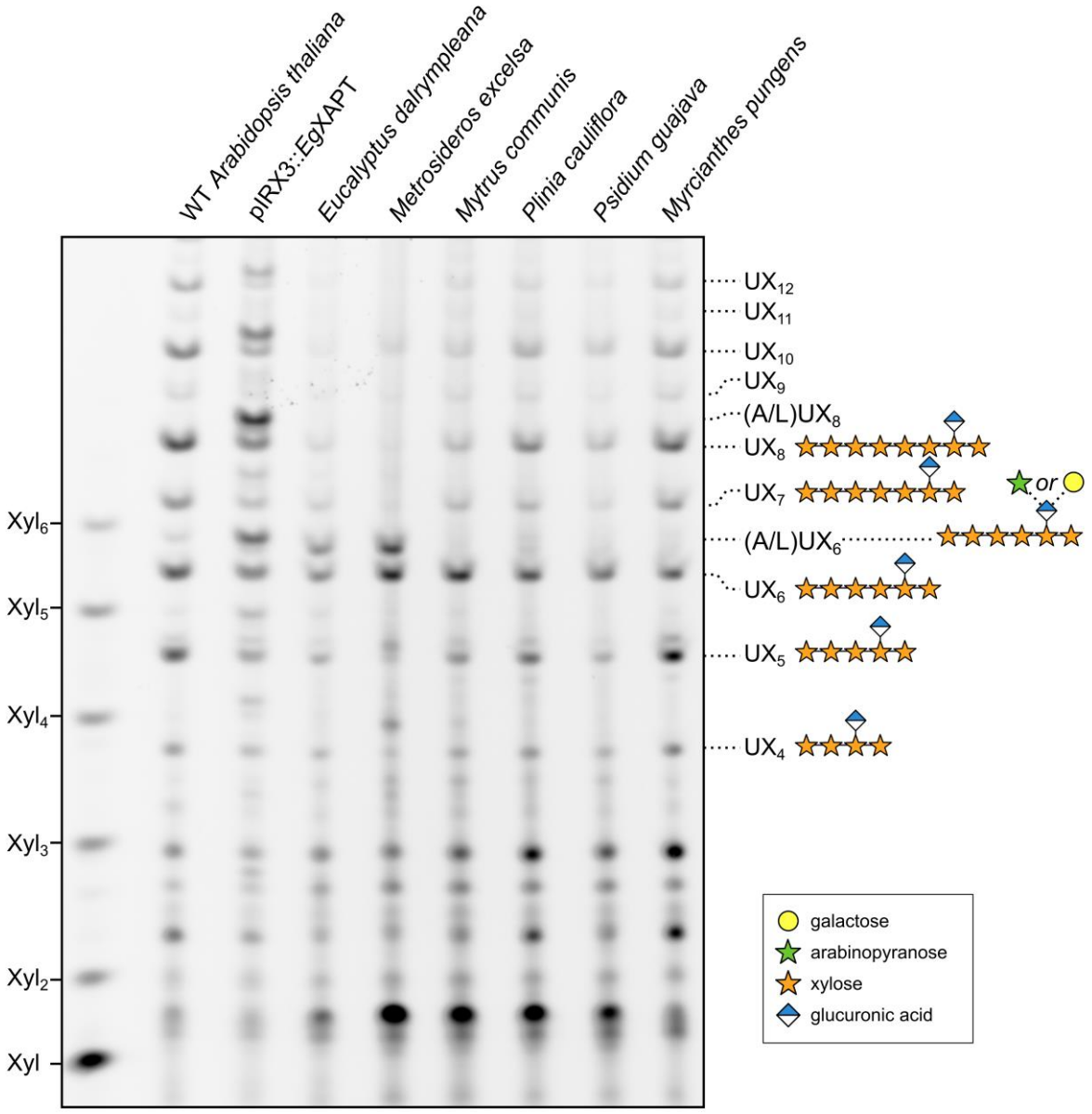


Figure 5.9 Xylan from phloem tissues in Myrtaceæ plants may exhibit a different spacing pattern compared with xylem tissues. Alkali-extracted hemicellulose from bottom stem AIR (in the case of Arabidopsis samples) or phloem/periderm AIR (in the case of Myrtaceæ samples) was digested with GH30 *endo*-xylanase from *Dickeya chrysanthemi* P860219. The products were derivatised with ANTS and separated by PACE.

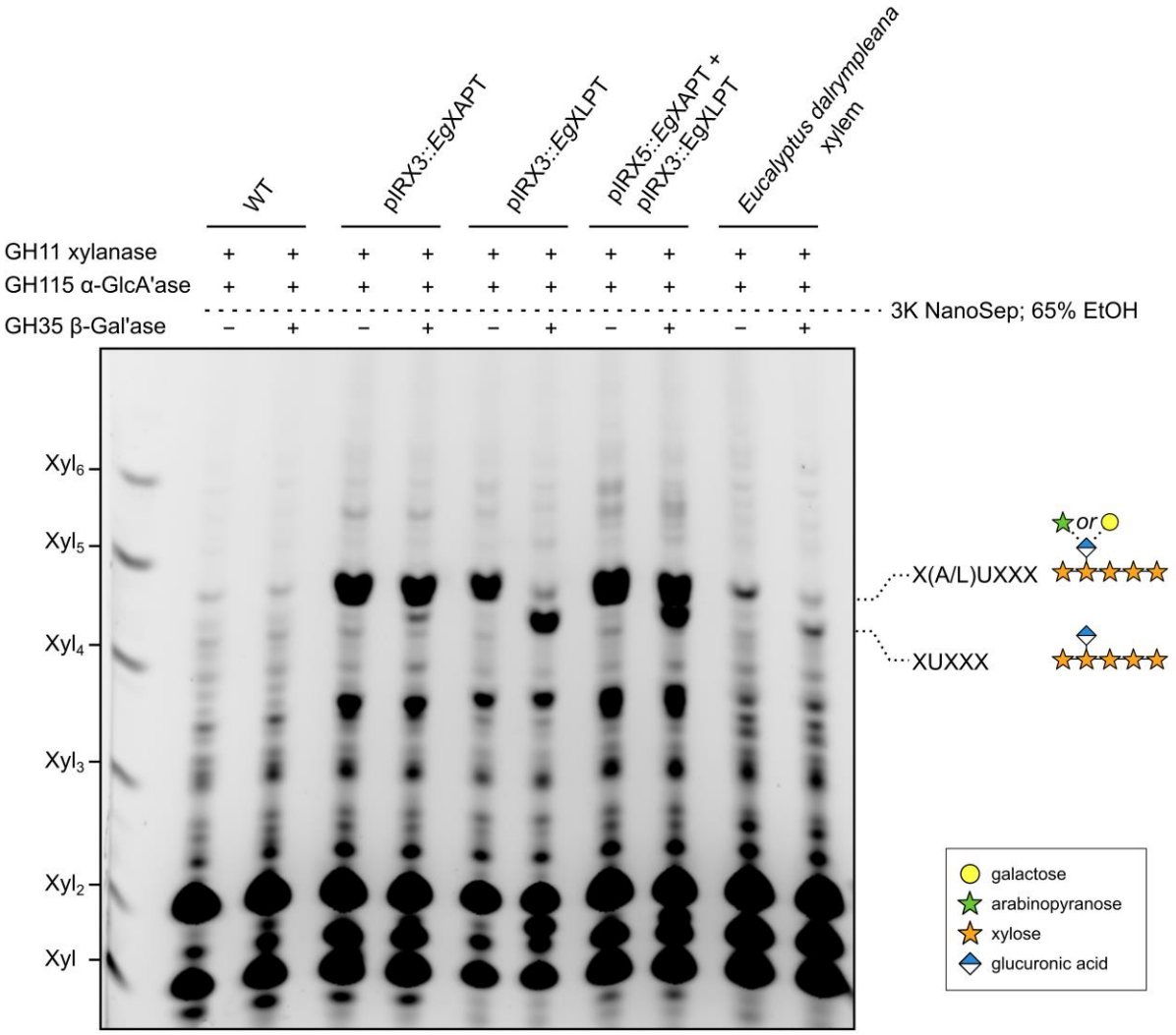


Figure 5.10 As in Arabidopsis, the presence of galactosylated GlcA residues in *Eucalyptus dalrympleana* xylan can be demonstrated by virtue of their sensitivity to β-galactosidase. Alkali-extracted hemicellulose from bottom stem AIR of Arabidopsis plants (WT and T_1 transgenic lines expressing *Eg*XAPT, *Eg*XLPT or *Eg*XAPT[A235G], all under the IRX3 promoter) or xylem tissue from xylem AIR of *Eucalyptus dalrympleana* was digested simultaneously with GH11 *endo*-xylanase and GH115 α-glucuronidase. The products were subsequently treated with GH35 β-galactosidase before ANTS derivatisation and analysis by PACE. The exposure used to produce this image is twice that used for the image shown in **Figure 5.7**.

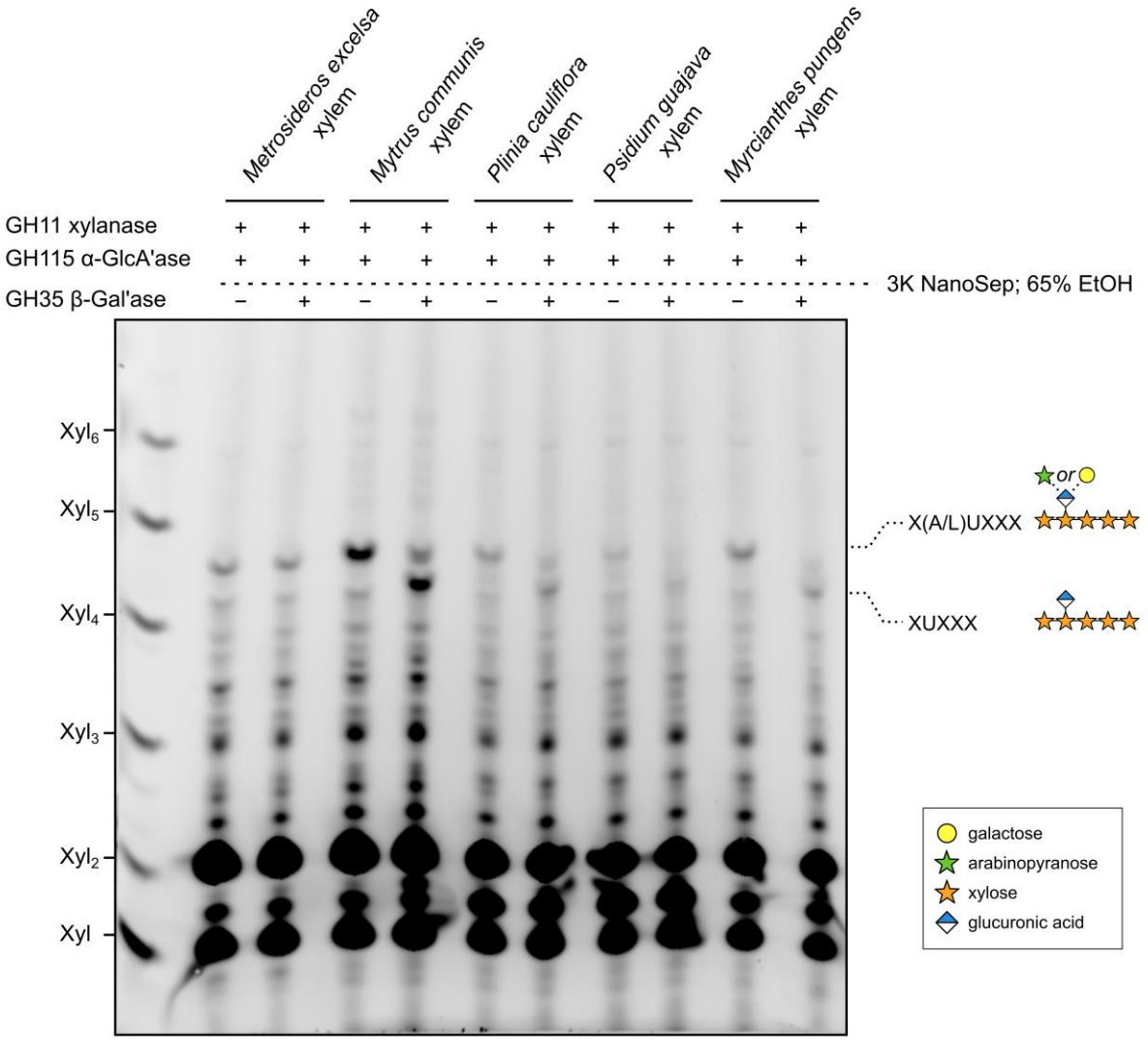


Figure 5.11 Many, but not all plants in the Myrtaceæ family appear to contain galactosylated GlcA sidechains on xylan in their xylem/pith tissues. Alkali-extracted hemicellulose from xylem/pith tissue AIR from the indicated species was digested simultaneously with GH11 *endo*-xylanase and GH115 α-glucuronidase. The products were subsequently treated with GH35 β-galactosidase before ANTS derivatisation and analysis by PACE.

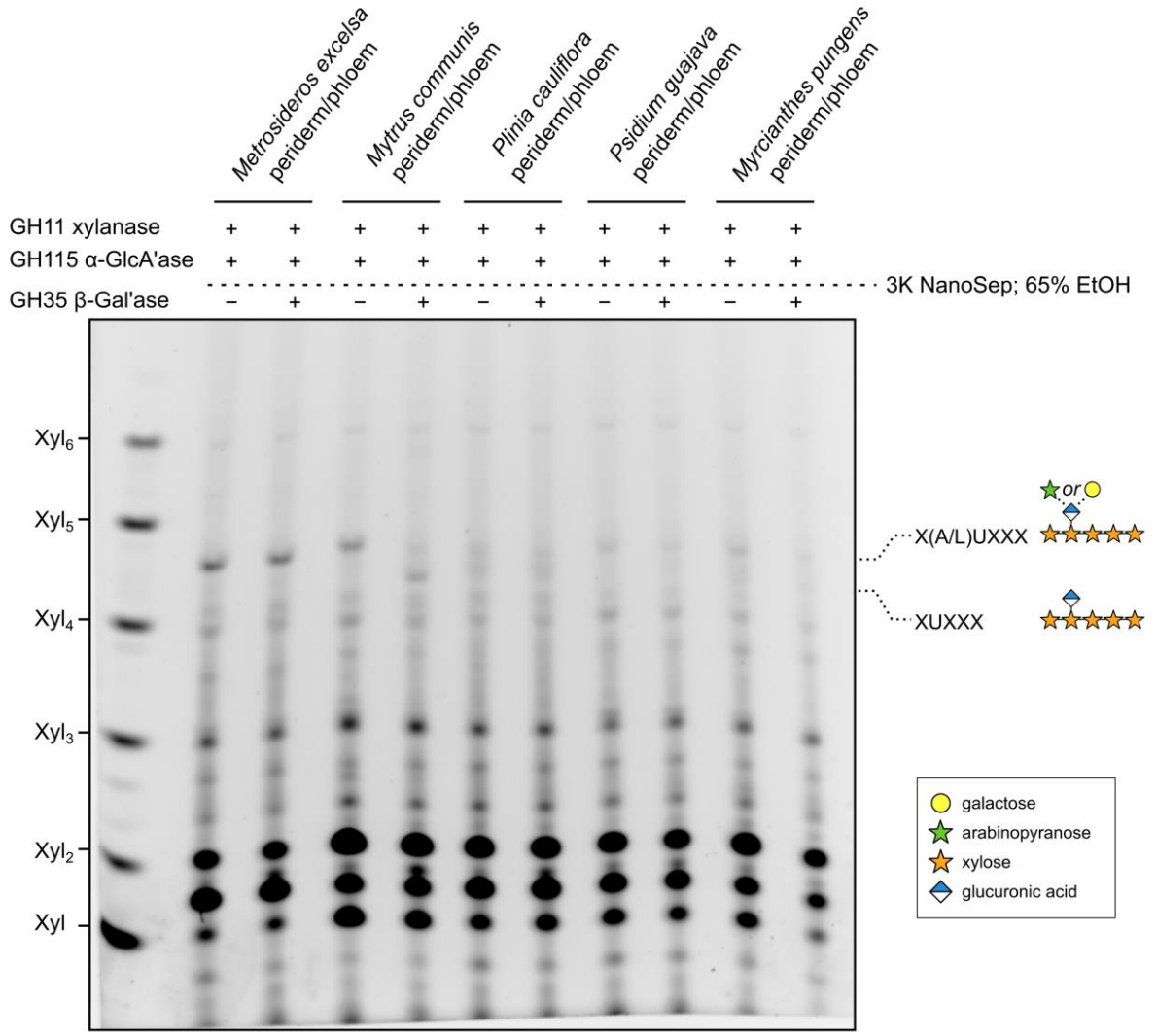


Figure 5.12 At least some plants in the Myrtaceæ family appear to contain galactosylated GlcA sidechains on xylan in their phloem/periderm tissues. Alkali-extracted hemicellulose from phloem/periderm tissue AIR from the listed species was digested simultaneously with GH11 *endo*-xylanase and GH115 α-glucuronidase. The products were subsequently treated with GH35 β-galactosidase before ANTS derivatisation and analysis by PACE. The amount of product loaded on the gel was twice that for the xylem/pith experiment shown in **Figure 5.11**.

5.2.6 XAPT and XLPT gene fragments can be amplified from the genomic DNA of many Myrtaceæ family members

Although I was able to detect the presence of XLPT-related genes in the published genomes of several Myrtaceæ species (see above), due to lack of material, I was not able to analyse the xylan structures in any of these particular species. Therefore, to facilitate correlation between the sequence and activity of XAPT and XLPT enzymes, I attempted to amplify *XAPT*- and *XLPT*-related coding sequences from genomic DNA extracts of *E. dalrympleana*, *M. excelsa*, *M. communis*, *P. cauliflora*, *P. guajava*, and *M. pungens* by PCR. Because introns in GT47-A coding sequences are reasonably rare (Tan *et al*, 2018; Xu *et al*, 2018; Wu *et al*, 2019), it was feasible to design primers for amplification of long stretches of coding sequence. Hence, I designed various primer pairs for *XAPT*/*XLPT* amplification based on conserved patches in published genomic sequences from the most closely related species. Through trial and improvement, I was able to amplify various fragments of *XAPT*- and *XLPT*-related coding sequences, which were then sequenced by the Biochemistry Department Sequencing Facility. The various primer pairs based on *XAPT*-related sequences were frequently able to amplify fragments of genomic DNA from all five species (**Table 5.1**). However, primers based on *XLPT* sequences were only successful in amplifying mid-length sequences from *E. dalrympleana* and *M. excelsa* (for which highly similar sequences from *E. grandis* and *M. polymorpha* were already available from published genomes). Attempts to amplify any similar sequences from the three Myrteæ-tribe species were unsuccessful except in the case of *M. pungens*, from which a short ~300 bp sequence (distinct from other sequences amplified with *XAPT*-based primers) could be amplified. These results suggest that plants in the Myrteæ tribe of the Myrtaceæ family exhibit a more divergent *XLPT*-type sequence (perhaps with introns that prevent simple PCR), or that this gene has been lost in these species and another similar enzyme has taken over the role of xylan galactosylation.

Table 5.1 Primers producing the largest PCR products from Myrtaceæ genomic DNA. Genomic DNA was prepared by CTAB extraction, and numerous PCR reactions were attempted with a number of primers designed to anneal to closely related sequences in available genomes. Only the primers producing the longest products are shown. Ed = *Eucalyptus dalrympleana*, Cc = *Corymbia citriodora*, Me = *Metrosideros excelsa*, Mc = *Myrtus communis*, Pc = *Plinia cauliflora*, Pg = *Psidium guajava* specimen 1 (fruit from Egypt), Pg2 = *Psidium guajava* specimen 2 (leaf from Brazil), Myp = *Myrcianthes pungens*. For *P. guajava* and *M. pungens*, sequencing of the products revealed the presence of more than one distinct amplicon, indicating the presence of two or more close homologues in the original genome, named here arbitrarily as *XAPT1* and *XAPT2*.

Gene	F primer sequence (5'→3')	lab code	R primer sequence (5'→3')	lab code	bp (apx.)
EdXAPT	CACTATCTCCATTCCAGAAACCC	2152	CATCACATTACACCAACACAATC	2153	1700
CcXAPT	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
MeXAPT	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
McXAPT	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
PcXAPT	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
PgXAPT1	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
PgXAPT2	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
Pg2XAPT1	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
Pg2XAPT2	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
MypXAPT1	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
MypXAPT2	TCATCGCCGAGATGATCTTCC	2241	AACCCCAGCTCCGTCG	2244	900
EdXLPT	CTCACTTCTCCTCCATAACCC	2147	CAAGCTAATTGGATCGATCAAGC	2150	1600
CcXLPT	CAACCAGTTCACCTCAGAGATGC	2237	GCCTCAAGTGCCACGTCG	2239	1000
MeXLPT	CTCTCTCCCTCTTCATCCTC	2235	ATGACTTTGGACCTCATCTTCTC	2240	1200
MypXLPT	GCCACCAACTATTTCACCTCAG	2291	GGTGGAAGTAGGAAGGGTAAGG	2294	300

To confirm the assignment of the amplified sequences to the *XAPT* or *XLPT* subclades, respectively (whose names I have chosen to reflect their phylogenetic grouping, and not necessarily their function), I wanted to construct a detailed phylogeny. I combined the

experimentally determined sequences with further *XAPT* and *XLPT* homologues from published Myrtales genomes (detected using the XAPT HMM as before), and constructed an alignment using MACSE (Ranwez *et al*, 2011, 2018). All sequences were then truncated to a region of 102 codon sites that was common to all sequences. A tree was then constructed under a codon substitution model using IQ-TREE (**Figure 5.13**). As expected, the sequences were separated into two distinct subclades, with the *XAPT* sequences grouped in one subclade, and the *XLPT* sequences grouped in the other. The branch lengths for *XLPT* sequences were longer than those for *XAPT*, suggesting that *XLPT* genes have undergone a greater amount of evolution since the initial divergence of *XAPT* and *XLPT*. By translating the aligned sequences into amino acid sequences, it was also possible to see the extent to which the (*Eg*XAPT *vs Eg*XLPT) Ala235 *vs* Gly212 distinction was conserved throughout the other Myrtaceæ sequences. As a matter of fact, although the Ala→Gly change was seen in *Eucalyptus* spp. XLPT sequences, it was *not* seen in XLPT sequences from the closely related *Corymbia citriodora* or any other species, confirming that the change is specific to the *Eucalyptus* genus (**Figure 5.13**). Interestingly though, one of the two XLPT-related sequences from *A. floribunda* exhibited an unusual Thr→Ile change at the preceding residue to the Ala/Gly residue of interest. These results suggest that residues in the vicinity of this position may experience different selection pressures in *XLPT* compared with *XAPT*.

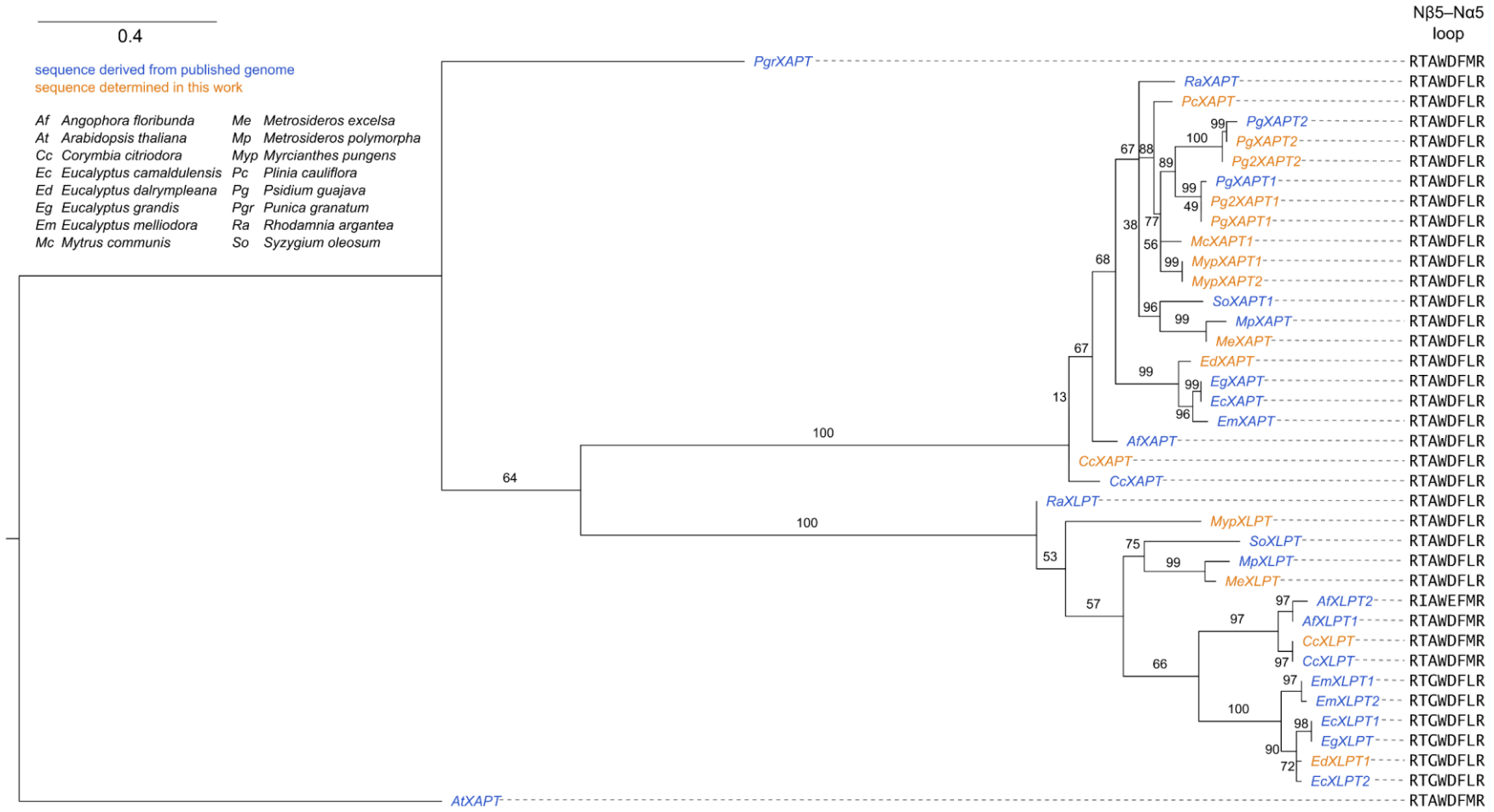


Figure 5.13 Detailed phylogeny of XAPT and XLPT coding sequences from Myrtaceæ-family genomes based on a well aligned segment comprising 102 codon sites. Thirty-eight sequences were aligned with MACSE before construction of a phylogeny using IQ-TREE. One thousand ultra-fast bootstrap replicates were carried out; branch labels display the percentage of replicates in which the corresponding split was present. For each taxon, the translated sequence of the predicted Nβ5–Nα5 loop is displayed.

5.3 Discussion

Plants possess a large number of GT47 glycosyltransferases involved in cell wall synthesis, and the substrate specificity of GT47 clade A is particularly diverse. Despite this diversity, all characterised GT47-A enzymes appear to share three aspects of their activity: they catalyse the formation of 1,2 glycosidic bonds, they invert the anomeric configuration of the donor sugar, and they transfer sugars to monosaccharide decorations on polysaccharide backbones. Therefore, although the types of monosaccharides involved can vary considerably within both the donors and acceptors of these enzymes, the overall shape of the product is essentially conserved. Therefore, it is easy to see how different activities could have arisen in this clade through subtle changes to the active site residues in these enzymes.

Here, I have presented a large-scale phylogeny of GT47-A protein sequences from across the plant kingdom, which demonstrates the considerable expansion of this clade from what appears to be only a handful of ancestral genes in the streptophyte algæ. The relationships between the subgroups are somewhat surprising and reveal that changes in substrate specificity have occurred multiple times in this family. For example, many of the groups identified here contain enzymes known to be involved in xyloglucan synthesis. However, the relationship between these groups is not especially close: the xyloglucan β-galactosyltransferase MUR3 appeared to be more closely related to the mannan β-galactosyltransferase MBGT1 than it was to XLT2, the other xyloglucan β-galactosyltransferase found in Arabidopsis, for instance. Furthermore, some functional orthologues appeared to belong to completely separate subclades: *Pp*XLT2 was not grouped in the *At*XLT2 subclade, for instance, but instead appeared to belong to a lower-plant-specific group that also included the xyloglucan arabinopyranosyltransferase *Pp*XDT. This suggests that convergent evolution has played a role in determining the functions of these enzymes.

The GT47-A phylogeny also revealed some interesting details regarding the evolution of cell wall-synthesising enzymes in monocots. Based on the branch lengths I observed, protein sequences from the Poaceæ family appear to have undergone significant evolution since their divergence from other commelinids. Poaceæ-family plants also appear not to possess a copy of XAPT. This is consistent with the fact that the Ara*p*-GlcA- disaccharide decoration has not been detected in grasses (Peña *et al*, 2016). Furthermore, XUT appears to have been lost from the Monocots entirely. These results are in concert with the fact that plants in the Poaceæ and

wider Poales family are well known to possess a distinct cell wall composition in comparison to eudicots and other monocots, with much lower amounts of xyloglucan and glucomannan (Carpita & Gibeaut, 1993; Smith & Harris, 1999; Burton & Fincher, 2012; Peña *et al*, 2016). The phylogeny also revealed the existence of a very large subclade of monocot-specific enzymes most closely related to MUR3 and MBGT1. So far, only one enzyme from this subclade has been characterised: *Sb*GT47_2 from *S. bicolor*, which acts as a galactosyltransferase to xyloglucan (Xu *et al*, 2018). Further investigation will be required to establish why such a significant clade expansion has arisen in these plants.

The donor substrates of XAPT and XLPT exhibit only a small difference in their chemical structure; hence, it is tempting to try to explain the difference in substrate specificity using sequence and structural data. I modelled the structure of *At*XAPT1 using the cryo-EM structure of EXTL3. Interestingly, in contrast to the Cα4 helix of the EXTL3 GT47 domain, which contains four turns, the Cα4 helix of *At*XAPT1 was predicted to contain three turns, like that of other characterised GT-Bs. This strengthens the argument that EXTL3 is an exception amongst GT47-family proteins, and that in other GT47s such as *At*XAPT1, the nucleotide sugar phosphates are bound at the N-terminus of this helix. Indeed, I was able to use this model to predict the residues that might determine nucleotide sugar specificity in GT47-As. I showed that *Eg*XAPT and *Eg*XLPT exhibit only a handful of differences near the predicted site of donor sugar binding; these changes could explain the difference in activity. However, by making a point mutant of *Eg*XAPT, I also showed that the most promising of these, Ala235→Gly in the Nβ5–Nα5 loop, is in itself insufficient to bring about a change in activity. Furthermore, by showing that galactosylated glucuronic acid decorations are more widespread in the Myrtaceæ family than this amino acid change is widespread amongst XLPT protein sequences, I have demonstrated that the amino acid change is also *unnecessary* for a change in activity. A logical conclusion might follow that the remaining, un-investigated amino acid changes at the N-terminus of Cα1 or in the Cβ2–Cα3 loop are responsible for the activity change; nevertheless, the determinants of GT substrate specificity have historically proven difficult to predict, and attempts to engineer substrate specificity in this manner have not always been so straightforward (Lairson *et al*, 2008; Chang *et al*, 2011)—accordingly, the difference in activity may have arisen from yet subtler, perhaps larger-scale perturbations to the nucleotide sugar binding site. Furthermore, although the Ala→Gly change may be neither necessary nor sufficient for galactosyltransferase activity, it may yet be *indicative* of differences in selection pressures between the two enzymes—perhaps in this case reflecting some finer optimisation of

the active site following neofunctionalization by other means. Consistent with this idea is the fact that Nβ5–Nα5-loop residues in other XAPTs (which are assumed to possess arabinopyranosyltransferase activity) appear to be highly conserved, implying that a selection pressure normally helps to maintain these residues as they are. However, on the flipside of this, if a particular XAPT homologue exhibits substantial changes to this loop, it could conceivably indicate a loss of selection pressure due to general loss of function or expression. Indeed, when I attempted to express one of the few XAPT genes with anomalous residues in the Nβ5–Nα5 loop in Arabidopsis bottom stem—FVE16326 from Rosaceæ member *Fragaria vesca*—I did not observe any alteration to the xylan structure (data not shown), suggesting that this enzyme is either not expressed or is inactive. It is likely that the better conserved FVE16329, close-by in the *F. vesca* genome, carries out the regular XAPT function in this plant. Hence, changes to the Nβ5–Nα5 loop residues could just as easily signal a loss of activity as well as they could signal a neofunctionalisation event.

In my phylogeny of the GT47-A clade, I found only one Arabidopsis enzyme (XAPT1) in group V. This implies that a single enzyme is responsible for synthesising substituted glucuronic acid xylan sidechains in Arabidopsis. In contrast, using more detailed phylogenies of group V sequences, I was able to show that most Myrtaceæ-family genomes contain both a XAPT-related and XLPT-related enzyme. Nevertheless, I was unable to detect a direct XLPT homologue in the *P. guajava* genome using BLAST and HMMER searches—and was also unable to amplify a XLPT-related sequence from *M. communis*, *P. cauliflora*, or *P. guajava* genomic DNA using PCR. The possibility that galactose could be transferred to glucuronic acid residues by some other enzyme in these plants therefore remains to be fully eliminated. The sequencing of further Myrteæ-tribe genomes and the characterisation of the remaining GT47-A enzymes will likely shed light on this subject.

To date, no specific function of the arabinopyranosylated branches of primary cell wall xylan has been identified, and the *xapt1* mutant exhibits no obvious growth phenotype (Yu *et al*, 2021b). Some insight into XAPT function may lie in the fact that some GH30 xylanases are unable to bind to and cleave at substituted glucuronic acid residues (Yu *et al*, 2021b). However, the enzymes that *are* able to do so are not able to distinguish galactosyl from arabinopyranosyl decorations (Yu *et al*, 2021b); hence, the selective pressures leading to the development of XLPT activity in the Myrtaceæ remain even less clear. Evtuguin *et al* (2003) have suggested that the galactosyl residues on *E. globulus* xylan could be covalently linked to a

rhamnoarabinogalactan. Such a linkage could presumably have important roles in cell wall architecture, and this hypothesis warrants further investigation.

In conclusion, with their high similarity, the XAPT and XLPT enzymes represent a useful model for investigating nucleotide sugar specificity in GT47-family glycosyltransferases and other GT-B enzymes. However, the difference between their activity may not be so easy to predict by rational structure-based approaches. Nevertheless, the work presented in this chapter provides the groundwork for more comprehensive evaluation of substrate specificity in these enzymes, as well as establishing a method for detecting the presence of galactoglucuronoxylan in uncharacterised plant cell walls. Furthermore, in the next chapter I will expand my focus to the whole of the GT47-A clade, and add more strength to the hypothesis that it might be possible to predict enzyme substrate specificity from the sequence of the Nβ5–Nα5 loop.

Chapter 6 : Nucleotide sugar specificity in the wider GT47-A clade

6.1 Introduction

In **Chapter 4**, I used the structure of EXTL3 to determine the fold of enzymes in the GT47 family. Then, in **Chapter 5**, I attempted to use this structural information to explain the difference in nucleotide sugar specificity between GT47-A family members XAPT and XLPT. More specifically, I investigated the idea that, within GT47-A glycosyltransferases, the amino acid composition of the Nβ5–Nα5 loop could bear some relation to donor substrate specificity. Although my results suggested that changes in this loop may not directly alter the enzyme activity, I was not able to rule out the possibility that changes in this loop could still be linked to, and therefore indicate, natural changes in activity. If this were true, it could permit the prediction of new GT47-A enzyme activities based on protein sequence.

Hence, I set out to study the correlation between sequence and substrate specificity across the entire GT47-A clade. However, rather than working on xylan (as XAPT and XLPT do), the majority of enzymes in this clade act in xyloglucan synthesis. As explained in *Section 1.5.3*, xyloglucan consists of a β-1,4 glucan backbone decorated with α1,6-linked xylosyl residues (often in a repeating pattern) that are themselves frequently substituted at the C2 hydroxyl by a second monosaccharide—usually β-galactose (which can be further substituted with fucose by FUT1). However, the identity of this second monosaccharide appears to exhibit variety across different tissues and species, with β-galacturonosyl, α-arabinofuranosyl, α-arabinopyranosyl, and β-xylosyl moieties having been reported in different plants and organs (of these, β-galacturonic acid and α-arabinopyranose may also be fucosylated) (Pauly & Keegstra, 2016). It is clear that GT47-A enzymes are responsible for adding many of these sugars: MUR3 and XLT2 act as galactosyltransferases in Arabidopsis, XUT is an Arabidopsis galacturonosyltransferase, *Sl*XST1 and *Sl*XST2 are arabinofuranosyltransferases from tomato, and *Pp*XDT is an arabinopyranosyltransferase from *P. patens* (**Figure 6.1a**) (Madson *et al*, 2003; Jensen *et al*, 2012; Peña *et al*, 2012; Schultink *et al*, 2013; Zhu *et al*, 2018). However, although Xyl-β1,2-Xyl-α1,6- xyloglucan sidechains have been identified in plants from the Ericales order (which include blueberries, cranberries, and the argan tree) (Ray *et al*, 2004; Hilz *et al*, 2007; Hotchkiss *et al*, 2015), a β-xylosyltransferase to the α1,6-linked xylose is yet to be identified. Nevertheless, given the structural resemblance of this sidechain towards other disaccharide xyloglucan decorations, it seems likely that the enzyme that forms this linkage is

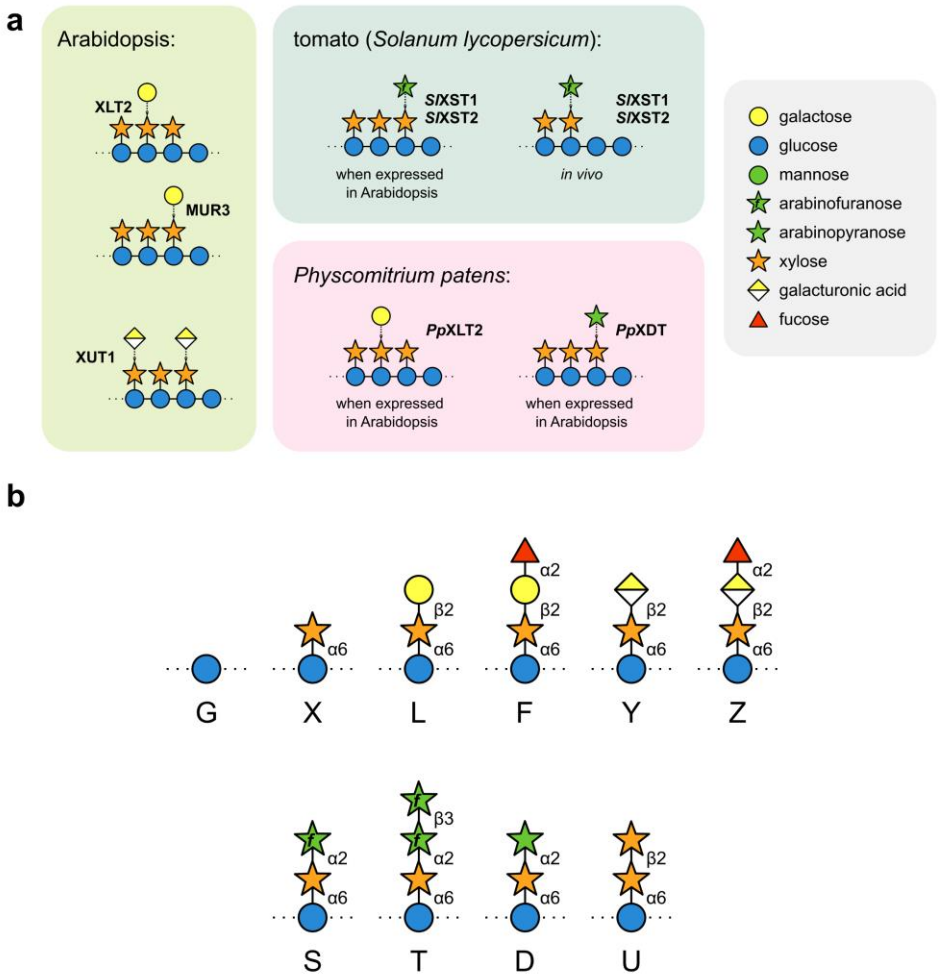


Figure 6.1 Xyloglucan-specific GT47-As and the sidechains they produce. a Characterised xyloglucan-specific GT47-A activities from *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Physcomitrium patens* (Madson *et al* 2013; Jensen *et al* 2012; Peña *et al* 2012; Schultink *et al* 2013; Zhu *et al* 2018). **b** Xyloglucan sidechain nomenclature.

also a member of GT47-A (Schultink *et al*, 2014). Therefore, if such an enzyme could be identified by virtue of its Nβ5–Nα5 sequence, the discovery would represent an excellent proof of concept for my proposed hypothesis.

Xyloglucan structures are usually described with reference to an established nomenclature that, for each backbone glucosyl residue, assigns a particular letter according to the type of sidechain attached. Undecorated glucose (Glc) is designated 'G', whereas glucose with a single α1,6-linked xylose decoration (Xyl-α1,6-Glc) is designated 'X' (**Figure 6.1b**) (Fry *et al*, 1993). Accordingly, plants that exhibit a decoration pattern in which every fourth glucose is undecorated are said to possess an 'XXXG' repeating core structure (Vincken *et al*, 1997). Furthermore, glucose with a galactosyl-xylose sidechain (Gal-β1,2-Xyl-α1,6-Glc) is assigned

'L', while the fucosylated version (Fuc-α1,2-Gal-β1,2-Xyl-α1,6-Glc) is assigned 'F'; similarly, glucose with an arabinofuranosyl-xylose (Ara*f*-α1,2-Xyl-α1,6-Glc), arabinopyranosyl-xylose (Ara*p*-α1,2-Xyl-α1,6-Glc), or xylosyl-xylose sidechain (Xyl-β1,2-Xyl-α1,6-Glc) is designated 'S', 'D', or 'U', respectively (Fry *et al*, 1993; Schultink *et al*, 2014).

In order to cleave between residues in the glucan backbone, canonical xyloglucanase enzymes require an unsubstituted glucose at the −1 position. Therefore, enzymatic xyloglucan digestion typically produces oligosaccharides with unsubstituted glucose at the reducing terminus—

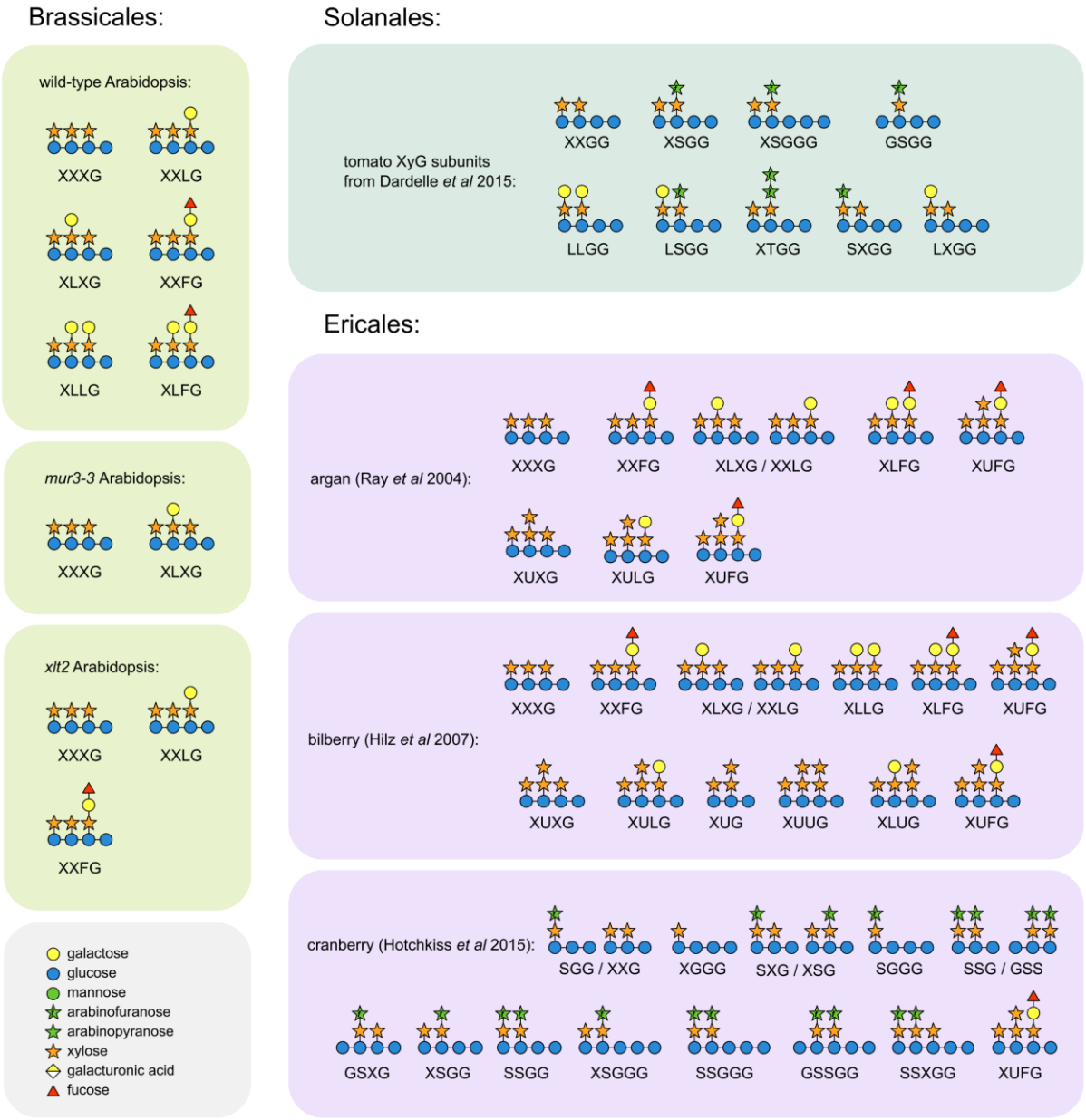


Figure 6.2 Xyloglucan subunits observed in different species and mutants. Oligosaccharide naming is applied according to the nomenclature of Frey *et al* (1993), as shown in **Figure 6.1**.

these structures have been termed xyloglucan 'subunits' (York *et al*, 1990; Pauly *et al*, 1999). For example, digestion of Arabidopsis xyloglucan produces the subunits XXXG, XLXG, XXFG, XLLG, and XLFG (**Figure 6.2**) (Pauly *et al*, 1999; Zabotina *et al*, 2008). Digestion of xyloglucan from Ericales plants, however, also produces β-xylosylated subunits such as XUXG, XUUG, XULG, and XUFG (Ray *et al*, 2004; Hilz *et al*, 2007; Hotchkiss *et al*, 2015). All of these subunits possess the XXXG structure at their core. However, this pattern is not universal: xyloglucans from plants in the Solanales family (which include tomato and potato), for instance, are thought to possess predominantly XXGG-based structures with arabinofuranosylated subunits such as XSGG, LSGG, and GSGGG (Schultink *et al*, 2014; Dardelle *et al*, 2015).

Arabidopsis GT47-A mutants have been useful in the characterisation of both endogenous and heterologous GT47-A xyloglucan-specific activities. Of these, the *mur3-3* mutant, which constitutes a T-DNA insertion-mediated knockout of *MUR3*, exhibits the most severe phenotype, with cabbage-like leaves and short stems (Tamura *et al*, 2005; Kong *et al*, 2015). The *mur3-1* mutant, however, constitutes a *MUR3* point mutant. Although it retains only a very small level of activity, the MUR3 protein continues to be expressed in this mutant—hence, *mur3-1* exhibits a less severe phenotype (Madson *et al*, 2003; Tamura *et al*, 2005; Jensen *et al*, 2012). Analysis of the xyloglucan structure in these mutants has revealed that MUR3 is responsible for adding galactose specifically to the *third* xylose in the XXXG repeat; hence, XXLG, XXFG, XLLG, and XLFG subunits are totally absent in the *mur3-3* mutant, leaving only XXXG and XLXG structures (Madson *et al*, 2003; Kong *et al*, 2015). In contrast, the only other known xyloglucan galactosyltransferase in Arabidopsis, XLT2, adds galactose specifically to the *second* xylose in the XXXG repeat, and *xlt2* knockout mutants exhibit a xyloglucan structure composed of XXXG, XXLG, and XXFG subunits, with a less severe phenotype compared with *mur3-3* (Jensen *et al*, 2012). The mildness of the *mur3-1* phenotype has been exploited to create an *xlt2 mur3-1* double mutant which, although lacking virtually all the normal galactosyl sidechains from its xyloglucan, possesses a less severe phenotype than *mur3-3* (Jensen *et al*, 2012; Kong *et al*, 2015). This mutant has provided a useful background for the characterisation of many heterologous xyloglucan-specific GT47-A activities (Schultink *et al*, 2013; Liu *et al*, 2015; Zhu *et al*, 2018). Additionally, the *mur3-3* mutant has been used to screen for heterologous xyloglucan galactosyltransferase activities by virtue of their ability to complement the strong phenotype (Lopes *et al*, 2010; Xu *et al*, 2018; Wang *et al*, 2020a). The typical methods used in these studies to characterise the modified xyloglucan

structures have been MALDI-TOF mass spectrometry, high performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD), and solution NMR.

In this chapter, I explore the relationship between the amino acid sequence of the Nβ5–Nα5 loop and the donor substrate specificity of enzymes in the GT47-A clade—and show that the two appear to be correlated across the family. I also demonstrate the activity of two previously uncharacterised GT47-A enzymes from cranberry and coffee—and show that their expression complements the phenotypes of the *mur3-3* and *xlt2 mur3-1* mutants. I report a new enzyme-based method for analysing xyloglucan structure, and, finally, I use it to demonstrate that these two GT47-A enzymes likely constitute a xyloglucan α-arabinofuranosyltransferase and a xyloglucan β-xylosyltransferase, respectively. The results in this chapter lay down a route to the discovery (and even creation) of novel GTs and polysaccharide structures, as well as shedding light on the diversity of hemicellulose structure across the plant kingdom.

6.2 Results

6.2.1 The identity of an amino acid triplet in the Nβ5–Nα5 loop correlates with nucleotide sugar specificity in GT47-A glycosyltransferases

To investigate whether the amino acid composition of the Nβ5–Nα5 loop might be linked to donor substrate specificity in GT47-A enzymes, I re-analysed the sequence alignment that was used to create the GT47-A tree shown in **Figure 5.2**. After removing sequences with indels in the predicted Nβ5–Nα5 loop (located with reference to the *At*XAPT1 sequence), I truncated all sequences to this very eight-residue portion. Next, I established several sequence clusters based on the original groupings in the phylogenetic tree, dividing them further when they contained multiple activities or distinct groups. For each cluster, aligned Nβ5–Nα5 loop sequences were submitted to WebLogo 3 (Crooks *et al*, 2004) to calculate the amino acid probability at each position. **Figure 6.3** shows the same tree as in **Figure 5.2**, but labelled with these clusters and the corresponding sequence logos. Due to the large number of angiosperm sequences included, some logos were likely biased towards angiosperm amino acid identities. Nevertheless, some remarkable conclusions could be drawn from these data. For instance, the last five amino acids of the loop (typically 'WDFRR') appeared to be strongly conserved both within and between most clusters (though the penultimate arginine showed weaker conservation). This suggests that these residues have an important role in enzyme activity or stability. However, interestingly, these residues did *not* differ between clusters containing enzymes with different nucleotide sugar specificities: the XUT cluster, the XST cluster, and the eudicot and monocot XLT2 clusters all exhibited a strong preference for WDFRR at these positions, for instance.

This demonstrates that the small amount of variation in these residues is unlikely to be able to explain the observed variation in donor specificity. In contrast, despite their strong conservation *within* each cluster, the first three amino acids of the Nβ5–Nα5 loop sequence were conspicuous in their pattern of variation *between* different clusters. In the various clusters containing the xyloglucan galactosyltransferases *At*XLT2, *Os*XLT2, *At*MUR3, and *Os*MUR3, for example (which, as discussed in **Chapter 5**, are not all closely related to each other), an 'RIT' motif was conserved at this location. Sequences in the MBGT1/GT15 cluster, which are assumed to represent glucomannan galactosyltransferases, also shared the first two amino acids, but replaced the threonine with an alanine or serine (RI[A/S]). Strikingly however, the small *At*XLT2-related cluster of sequences containing the xyloglucan arabinofuranosyltransferases *Sl*XST1 and *Sl*XST2 did not exhibit an RIT motif, but rather an R**L**T motif, with isoleucine totally absent from the second position. Furthermore, all clusters thought to contain enzymes with non-galactosyltransferase activities exhibited different motifs from the galactosyltransferase clusters: KI[S/T] for the XUT galacturonosyltransferase cluster and RT[A/S] for the angiosperm XAPT arabinopyranosyltransferase cluster (as explored in **Chapter 5**), for instance. The RIT motif was not seen in any of these non-galactosyltransferase enzymes. Similarly, *Pp*XLT2 galactosyltransferase and *Pp*XDT arabinopyranosyltransferase (from the 'lower plant-specific cluster') showed differences in the same region: the individual Nβ5–Nα5 loop sequences were RI**V**WDFVR and RI**F**WDHNR, respectively. These data support the notion that the donor sugar specificity of GT47-A enzymes could be somehow related to the sequence of the Nβ5–Nα5 loop.

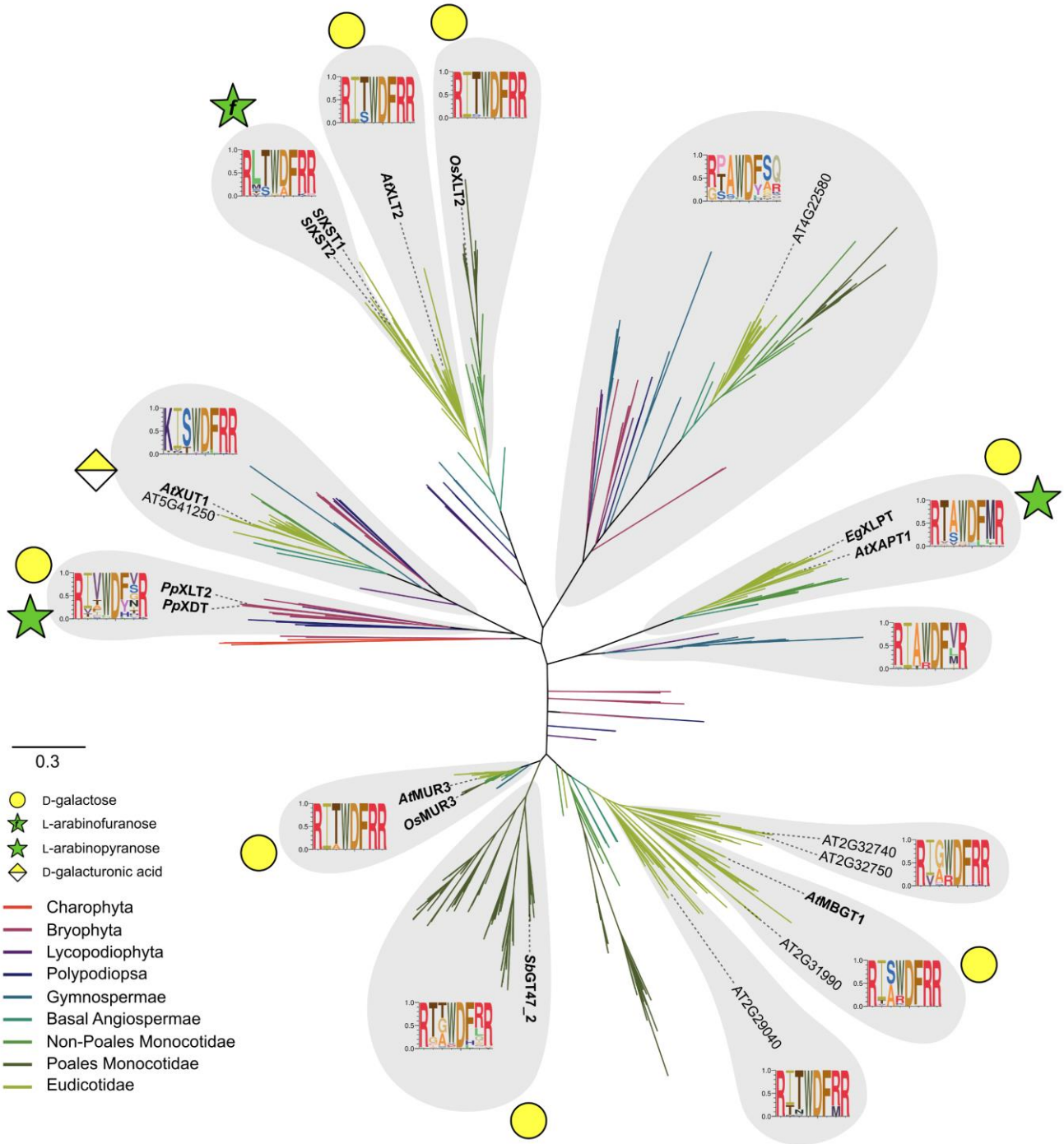


Figure 6.3 Phylogeny from Chapter 5, annotated with experimental clusters and their corresponding sequence logos. Clusters were rationally determined based on the pre-existing phylogenetic clades and the different activities they contained. Nβ5–Nα5 loop sequences were extracted from the overall alignment; sequence logos, which in this case show amino acid probabilities at each site, were created with WebLogo 3 (http://weblogo.threeplusone.com/). The nucleotide sugar specificities of various characterised enzymes throughout the tree are labelled using monosaccharide symbols.

6.2.2 Cranberry and coffee genomes encode XLT2/XST homologues with unusual residues in the predicted Nβ5–Nα5 loop

Because they exhibit β-xylosylated xyloglucan, plants in the Ericales family have been proposed to possess a GT47-A xyloglucan xylosyltransferase (as discussed above) (Schultink *et al*, 2014). I wanted to search for such a xylosyltransferase on the basis that it might exhibit unusual residues in its Nβ5–Nα5 loop. However, the large GT47-A tree did not include any Ericales sequences except for *Actinidia chinensis* (kiwi), for which the presence of the β-xylosylated xyloglucan structure has not been tested. Hence, I downloaded genomic data for the Ericales plants *Argania spinosa* (argan), *Camellia sinensis* (tea plant), *Rhododendron williamsianum*, *Vaccinium corymbosum* (blueberry), and *Vaccinium macrocarpon* (cranberry). For those species with proteome models, I collected GT47-A sequences using an Arabidopsis GT47-A HMM; otherwise, I employed TBLASTN, using the Arabidopsis GT47-A sequences as a multi-sequence query. The extracted sequences were combined with GT47-A sequences from Arabidopsis, rice, tomato, and kiwi before alignment. After truncation to the predicted GT47 domain, a phylogeny was constructed using RAxML. Examination of the resultant tree (**Figure 6.4**) revealed that the sequences could be sorted into the same six subclades as previously. Remarkably, however, *V. macrocarpon*, *V. corymbosum*, and *R. williamsianum* all exhibited an expansion in copy number for sequences in the XST subclade. Therefore, I examined the XLT2/XST subclade and its sequences in more detail (**Figure 6.5**). Intriguingly, many of the Ericales sequences exhibited altered residues at the amino acid triplet of interest in the Nβ5–Nα5 loop, with several sequences exhibiting an 'RMT' motif in place of the RLT triplet seen in closely related XST enzymes. I hypothesised that these enzymes might possess a novel nucleotide sugar specificity, and that, more specifically, they could harbour β-xylosyltransferase activity.

Furthermore, when I looked more closely at the XST subclade in the larger GT47-A tree, I noticed that sequences belonging to *Coffea canephora* (robusta coffee;

Table 6.1 Nβ5–Nα5 loop sequences from XLT2/XST homologues in *Coffea canephora*.

Protein	Nβ5–Nα5 loop
***At*XLT2 orthologues**	
Cc02_g03750	RITWDFRR
Solyc02g092840.1 orthologues	
Cc07_g06510	RITWDFRR
***Sl*XST1/2 orthologues**	
Cc05_g01260	RLTWDFRR
***Sl*XST1/2 paralogues**	
Cc01_g03040	RTSWDFKR
Cc07_g06540	RSSWDFRR
Cc07_g06550	RSSWDFRR
Cc07_g06570	RVSGDFRP
Cc07_g06590	RMSWDFRR

Gentianales) exhibited an expansion in copy number and exhibited altered residues in the Nβ5–Nα5 loop. In fact, the Nβ5–Nα5 loop triplet differed substantially even amongst these close homologues from *C. canephora*, which appear to be situated close to one another in the *C. canephora* genome (**Table 6.1**). Therefore, I also hypothesised that one or more of these sequences might encode an enzyme with a novel activity.

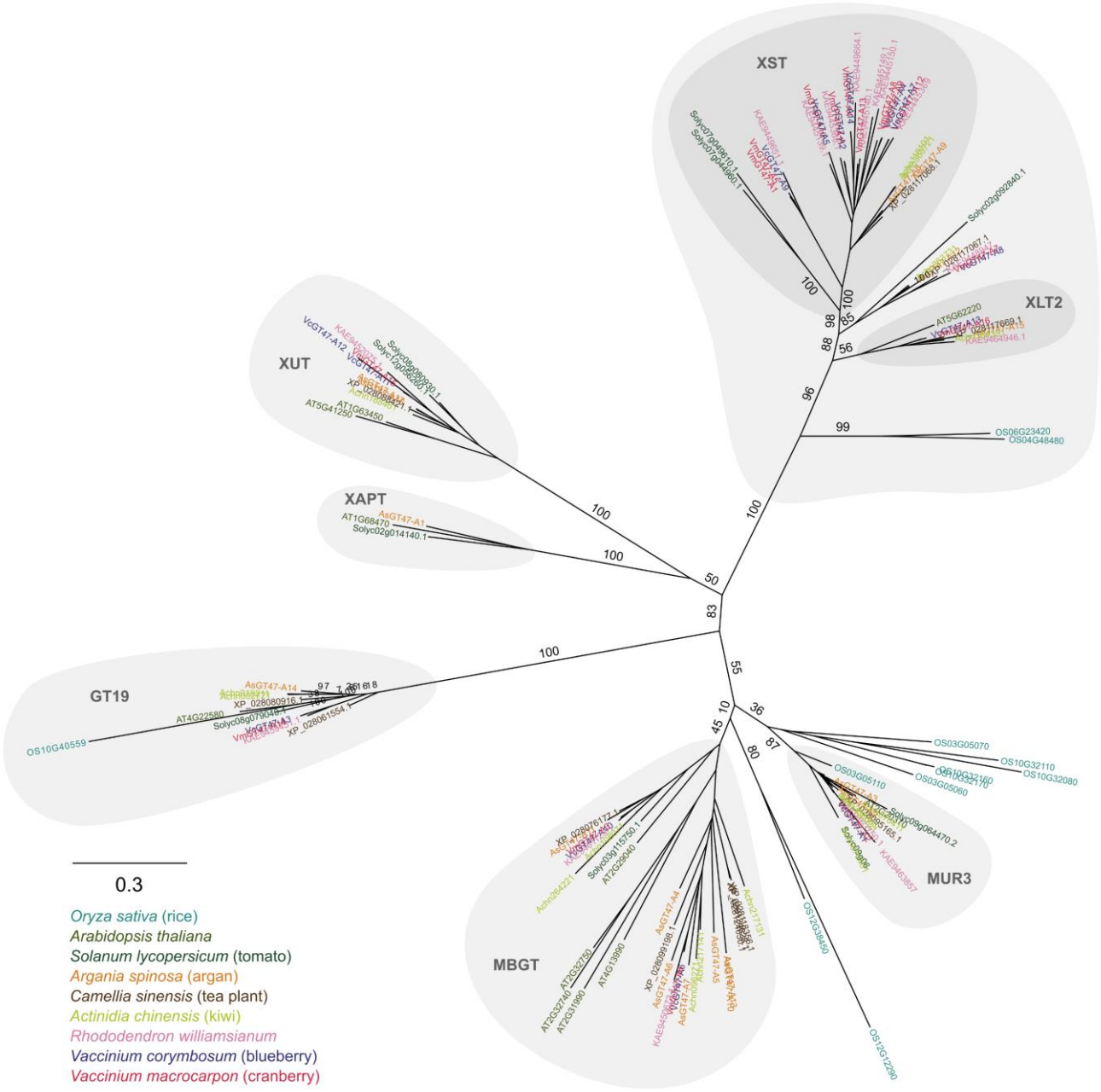


Figure 6.4 Phylogeny of GT47-A sequences from plants in the Ericales order. GT47-A protein sequences were obtained from six Ericales genomes by HMM or TBLASTN searches. These sequences were aligned with similar sequences from Arabidopsis, tomato, and rice. After truncation of the sequences to their GT47 domain, a phylogeny was constructed using RAxML. Branch labels (shown only for major branches) show the number of times each split was replicated in 100 ultra-fast bootstrap replicates.

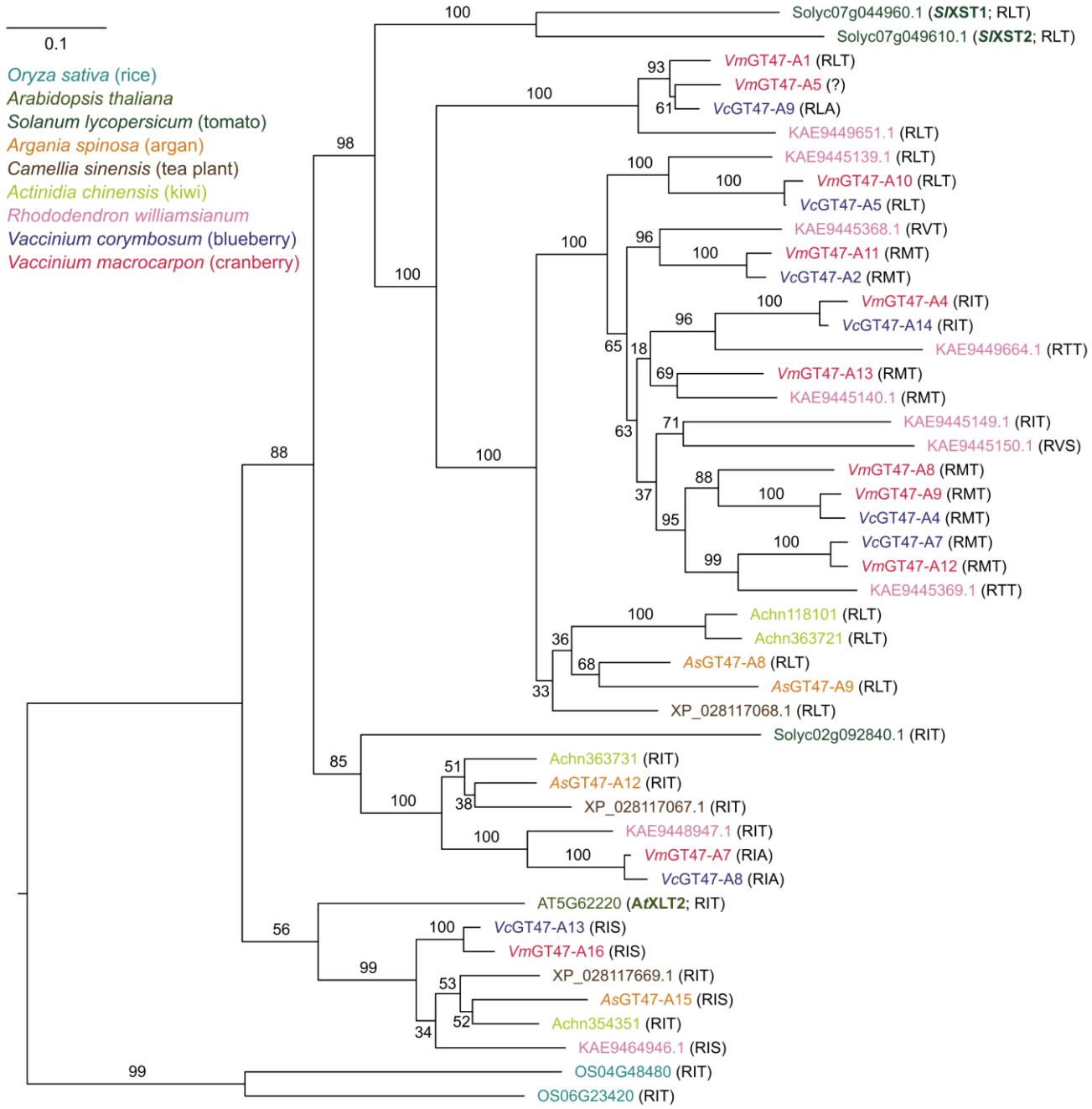


Figure 6.5 XLT2/XST subtree from the previous phylogeny of Ericales GT47-A sequences. See previous figure for methodology.

Finally, I noticed that plants from the Caryophyllales order, which comprised *Amaranthus hypochondriacus*, *Chenopodium quinoa*, and *Beta vulgaris* in the large GT47-A tree, appeared to possess an extra homologue of XUT1 with non-canonical residues in the Nβ5–Nα5 loop (data not shown). Similarly, I speculated that enzymes in this small subclade might possess an activity that differs from the galacturonosyltransferase activity of Arabidopsis XUT1.

6.2.3 Expression of Cc07_g06550 and Vm*GT47-A12 rescues the phenotypes of* mur3-3 *and* xlt2 mur3-1 *Arabidopsis mutants*

Because of their interesting amino acid motifs, I wanted to characterise the activity of some of these enzymes. However, when I previously made transgenic Arabidopsis lines expressing FVE16326, a XAPT-related GT47-A with an anomalous Nβ5–Nα5 loop sequence, I did not detect any changes in cell wall polysaccharide structure. This presented the possibility that similar experiments on other anomalous GT47-A enzymes might also be fruitless. Hence, I wanted to execute a screening procedure wherein xyloglucan-modifying activity could be detected prior to cell wall analysis. Therefore, I chose to express the enzymes of interest in the *mur3-3* mutant, whose strong phenotype makes it possible to detect xyloglucan-specific activity by virtue of phenotypic complementation (be it partial or full).

Amongst the Ericales xylosyltransferase candidates, I particularly wanted to investigate the enzymes from *V. macrocarpon* or *A. spinosa*, as these species have already been shown to possess the β-xylose-containing 'U' sidechain. However, using TMHMM, a transmembrane domain could be confidently predicted in only one of the especially interesting RMT-motif-containing *V. macrocarpon* sequences: *Vm*GT47-A12. I therefore chose to investigate this enzyme in particular. However, the protein sequence of *Vm*GT47-A12 appeared to contain several variable repeats in its stem domain, making it rather long. Since the very closely related *Vc*GT47-A7 from *V. corymbosum* had fewer repeats, I decided to remove two from the *Vm*GT47-A12 sequence to facilitate cloning and expression; hence, nucleotides corresponding to amino acids 167–193 were omitted from the expression construct (**Figure 6.6**). For the *C. canephora* sequences, I chose to express Cc07_g06550 and Cc07_g06570 without alteration. Similarly, for the Caryophyllales enzymes, I elected to express Bv7_174350_exki from *B. vulgaris* without modification.

Accordingly, I generated transgenic *mur3-3* Arabidopsis plants expressing *Vm*GT47-A12, Cc07_g06550, Cc07_g06570, or Bv7_174350_exki under the control of the promoter of XXT2 (one of the α-xylosyltransferases that synthesises xyloglucan during primary cell wall

deposition). To inspect for complementation of the mutant phenotype, I examined the stature and overall morphology of six-week-old T_1-generation plants. Plants expressing Cc07_g06570 or Bv7_174350_exki, which had short stature, curled, cabbage-like leaves, and short siliques, showed no substantial difference in phenotype to untransformed *mur3-3* plants (**Figure 6.7**). In contrast, expression of *Vm*GT47-A12 or Cc07_g06550 appeared to result in a consistent complementation effect. Plants expressing Cc07_g06550 invariably exhibited a phenotype close to that of wild-type Arabidopsis (WT), whereas the phenotypes of *Vm*GT47-A12-expressing lines, though somewhat variable, were always intermediate between those of *mur3-3* and wild type. These results suggest that *Vm*GT47-A12 and Cc07_g06550 possess some sort of xyloglucan-specific activity whose product compensates for the lack of 'L' and 'F' sidechains in the *mur3-3* mutant.

```
VcGT47-A7     MANTTFPPTFGQPSTHLQGLKKQPNKSPPLKSTVKSFLPALTPTGRTLVFIVILFNVFLV (60)
VmGT47-A12    MANTSFLPTFGHPSTNLQGLKKQPNKSPPLNSTLKSFLHSLTPTRRTLVIIVILFNVFLV (60)
              ****:* ****:***:***************:**:**** :**** ****:**********

VcGT47-A7     LYLTRTHILSPSESPRASPEFLPGESRIVRNISSQTYGGYSSLQNKDVLNISSQAY---D (117)
VmGT47-A12    LYLTRTHILSPSESPLASPEFLPGESRIVRNISSQTYSGSGSLQNKGVLNISSQAYELHD (120)
              *************** **********************.* .*****.*********   *

VcGT47-A7     GSPDNSLSDISSQT---------------------------------------------- (131)
VmGT47-A12    SSPDYDAPDISSQTYGRLRNKGVVDIESQNHGSLANKGVVNISSETYGSLANKEVLNISS (180)
              .*** . .******

VcGT47-A7     -------------YGSLANKEVMNISSEANKGVENISSGTYGGLANKDGLNISSETYGGL (178)
VmGT47-A12    EANKGIVNISSESYGSLANKEVLNISSEANKGVENISSGPYGGLANKDGLNISSETFGGL (240)
                          *********:***************.***************:***

VcGT47-A7     PNKSCDSGR... (187)
VmGT47-A12    PNKSCDSGR... (249)
              *********
```

Figure 6.6 Residues 167–193 of the *Vm*GT47-A12 coding sequence were omitted in construct design. *Vm*GT47-A12 was aligned to its close homologue *Vc*GT47-A7 using MUSCLE (only the CTS domain residues are shown). Underlined residues were omitted when designing the expression construct.

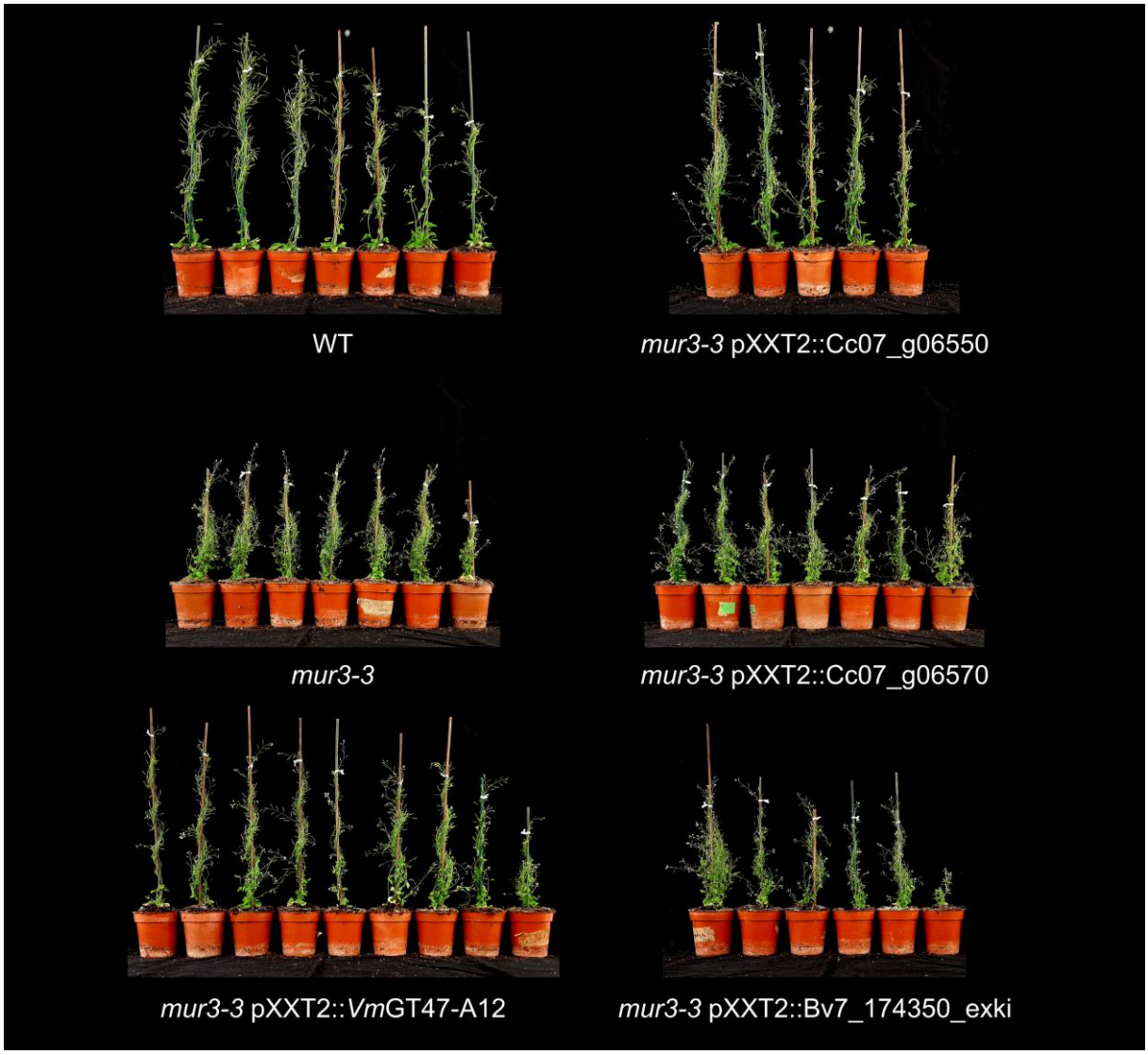


Figure 6.7 Growth phenotypes of *mur3-3* transgenic lines (T_1 generation) after six weeks of growth. *mur3-3* Arabidopsis plants were transformed by floral dip to introduce constructs for expression of *Vm*GT47-A12, Cc07_g06550, Cc07_g06570, or Bv7_174350_exki under the control of the XXT2 promoter. For transgenic plants, each plant represents an individual transformant. Plants were placed in order of height for photography.

Consequently, I selected *Vm*GT47-A12 and Cc07_g06550 for further analyses. To facilitate detailed characterisation of their activities, I chose to express these enzymes in the *xlt2 mur3-1* mutant, which possesses a xyloglucan structure composed almost entirely of XXXG subunits (Jensen *et al*, 2012) (and hence provides a maximal number of acceptor sites for GT47 activity). I generated transgenic *xlt2 mur3-1* Arabidopsis plants expressing *Vm*GT47-A12, Cc07_g06550, or the previously characterised arabinofuranosyltransferase *Sl*XST1 under the control of the CESA3 promoter, which induces very strong expression in tissues undergoing primary cell wall synthesis. As before, I examined the phenotypes of six-week-old T_1 plants. Interestingly, the same pattern of phenotypes was replicated: whereas over-expression of either *Sl*XST1 or Cc07_g06550 restored the level of growth to that of wild-type plants (or perhaps even higher), expression of *Vm*GT47-A12 resulted only in partial complementation, in spite of the use of the strong CESA3 promoter (**Figure 6.8**). These results suggest that, whereas *Sl*XST1 and Cc07_g06550 exhibit similar activities/levels of activity in these plants, *Vm*GT47-A12 may differ in its activity or level of activity.

6.2.4 Cc07_g06550 and Vm*GT47-A12 encode xyloglucan-specific pentosyltransferases*

To confirm that *Vm*GT47-A12 and Cc07_g06550 are able to modify xyloglucan structure in Arabidopsis, I wanted to analyse the xyloglucan structure in my *mur3-3* transgenic lines. To that end, I used alkali to extract hemicellulose from upper-stem or leaf AIR of the *mur3-3* plants expressing Cc07_g06550 or *Vm*GT47-A12, as well as from AIR of wild-type, *fut1*, and *mur3-3* Arabidopsis plants (provided by Dr Li Yu). These hemicellulose preparations were then treated with GH5 *endo*-xyloglucanase to produce a mixture of oligosaccharides with undecorated glucose at the reducing terminus. Subsequently, an aliquot of each sample was derivatised with the ANTS fluorophore and separated by PACE. PACE analysis revealed that, as expected, the xyloglucanase products from wild-type plants constituted XXXG, XXLG, XXFG, XLLG, and XLFG (**Figure 6.9**; assignments were made with reference to previously documented migrations (Yu *et al*, 2021a)). Furthermore, the xyloglucan subunits released from *fut1* material could be identified as XXXG, XXLG, and XLLG, whereas those from *mur3-3* constituted XXXG and XLXG. The xyloglucanase products from *mur3-3* plants expressing *Vm*GT47-A12 appeared to be very similar to those from the untransformed *mur3-3* plants, although an extremely faint band was just perceptible above the position of the XXXG band. In contrast, the xyloglucanase products from *mur3-3* plants expressing Cc07_g06550 were clearly different, exhibiting a new band that ran below XLXG—at a similar position to that of

XXLG. These results demonstrate that, whereas Cc07_g06550 has a clear ability to modify Arabidopsis xyloglucan structure, *Vm*GT47-A12[Δ167–193] may possess only a weak xyloglucan-modifying activity, if any at all.

Figure 6.8 Growth phenotypes of *xlt2 mur3-1* transgenic lines (T_1 generation) after six weeks of growth. *xlt2 mur3-1* Arabidopsis plants were transformed by floral dip to introduce constructs for expression of *Sl*XST1, Cc07_g06550, or *Vm*GT47-A12 under the control of the CESA3 promoter. For the transgenic plants shown in the lower panels, each plant represents an independent transformant. The top panel shows a line-up of the representative plants from each genotype, photographed side by side.

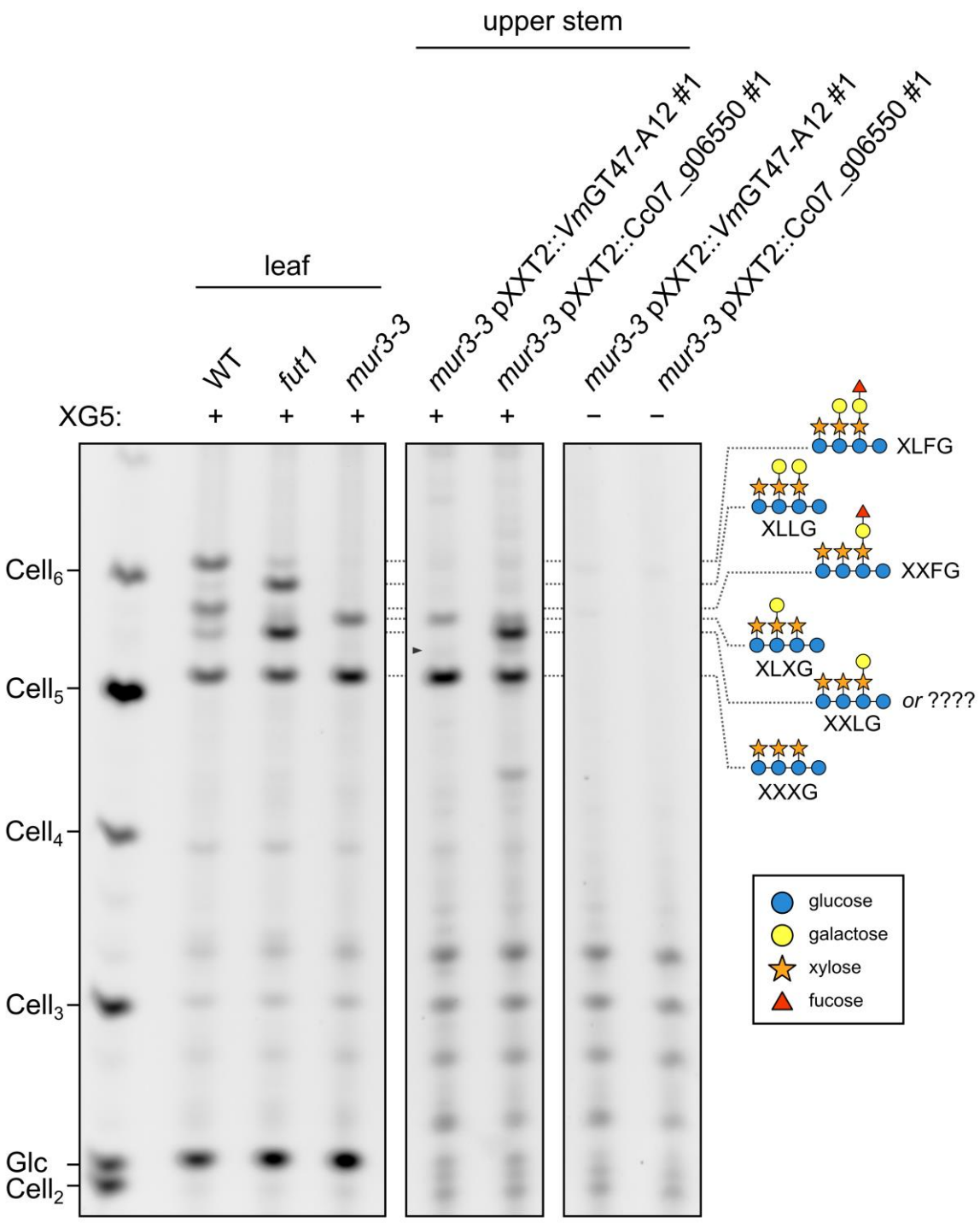


Figure 6.9 Expression of Cc07_g06550 alters xyloglucan structures in *mur3-3* plants. AIR from leaves or upper stem of the indicated Arabidopsis genotypes was treated with alkali to extract hemicellulose for subsequent digestion with the xyloglucanase XG5. The products were derivatised with ANTS and separated by PACE. Cello-oligosaccharides were used as standards. This figure shows a single gel after digital cropping of two irrelevant lanes. A grey arrowhead marks the position of an extremely faint band present in the digest. For each transgene construct, only one independent line was analysed (line #1 in both cases).

Novel xyloglucan structures have previously been analysed by oligosaccharide mass profiling (OLIMP) (Günl *et al*, 2010)—that is, by analysing the masses of xyloglucanase products by MALDI-TOF mass spectrometry (MS). This technique can reveal the number of hexosyl, pentosyl, and deoxyhexosyl residues within each xyloglucan subunit. Accordingly, MALDI-TOF MS was used to further analyse the structures of these xyloglucanase products—this time not only from the *Vm*GT47-A12 and Cc07_g06550-expressing lines, but also those expressing Cc07_g06570 and Bv7_174350_exki. For this experiment, hemicellulose was extracted from leaf AIR. I prepared xyloglucanase products for mass spectrometry; Dr Li Yu operated the mass spectrometer. As expected, the products from wild-type xyloglucan were apparent as ions with m/z ratios 1085.4, 1247.6, 1393.6, 1409.6, and 1555.6, which were interpreted as sodium adducts of the oligosaccharides XXXG, XLXG/XXLG, XXFG, XLLG, and XLFG, respectively (**Figure 6.10**). Furthermore, analysis of the *mur3-3* xyloglucanase products revealed the presence of only the first two of these ions, *i.e.* sodiated XXXG and XLXG. As anticipated, the spectra for lines expressing Cc07_g06570 or Bv7_174350_exki, which showed no complementation effect, were very similar to that of *mur3-3*. However, strangely, the XXXG ion was present predominantly as an M+1 peak in these spectra, with an m/z ratio of 1086.5–6 (this phenomenon could not be explained, and appears to affect various ions in various spectra at random). Nevertheless, this suggests that the xyloglucan in these plants also exhibits a structure with only XXXG and XLXG subunits. In contrast, xyloglucanase digests from lines expressing either *Vm*GT47-A12 or Cc07_g06550 produced an ion with an m/z ratio of 1217.5–6, which corresponds to an oligosaccharide containing four hexoses and four pentoses (H_4P_4). Based on the typical structure of Arabidopsis xyloglucanase products and the likely activity of these enzymes, this product plausibly constitutes an oligosaccharide with the structure X☆XG or XX☆G, where '☆' represents Pent-1,2-Xyl-α1,6-Glc—*i.e.* the S, U, or D sidechain. Interestingly, whereas the intensity of this peak was relatively strong in all three Cc07_g06550-expressing lines, this ion could only be detected for two of the three *Vm*GT47-A12-expressing lines, and furthermore, was very weak in these two lines. On the other hand, an ion with m/z ratio 1379.6, albeit with very weak intensity, was also detected for the same two *Vm*GT47-A12-expressing lines. This ion likely corresponds to a product with structure XL☆G. If present in the products from Cc07_g06550-expressing lines, it was barely detectable. As a whole, these results are consistent with the conclusions derived from PACE analysis, and indicate that Cc07_g06550 and *Vm*GT47-A12 are likely pentosyltransferases.

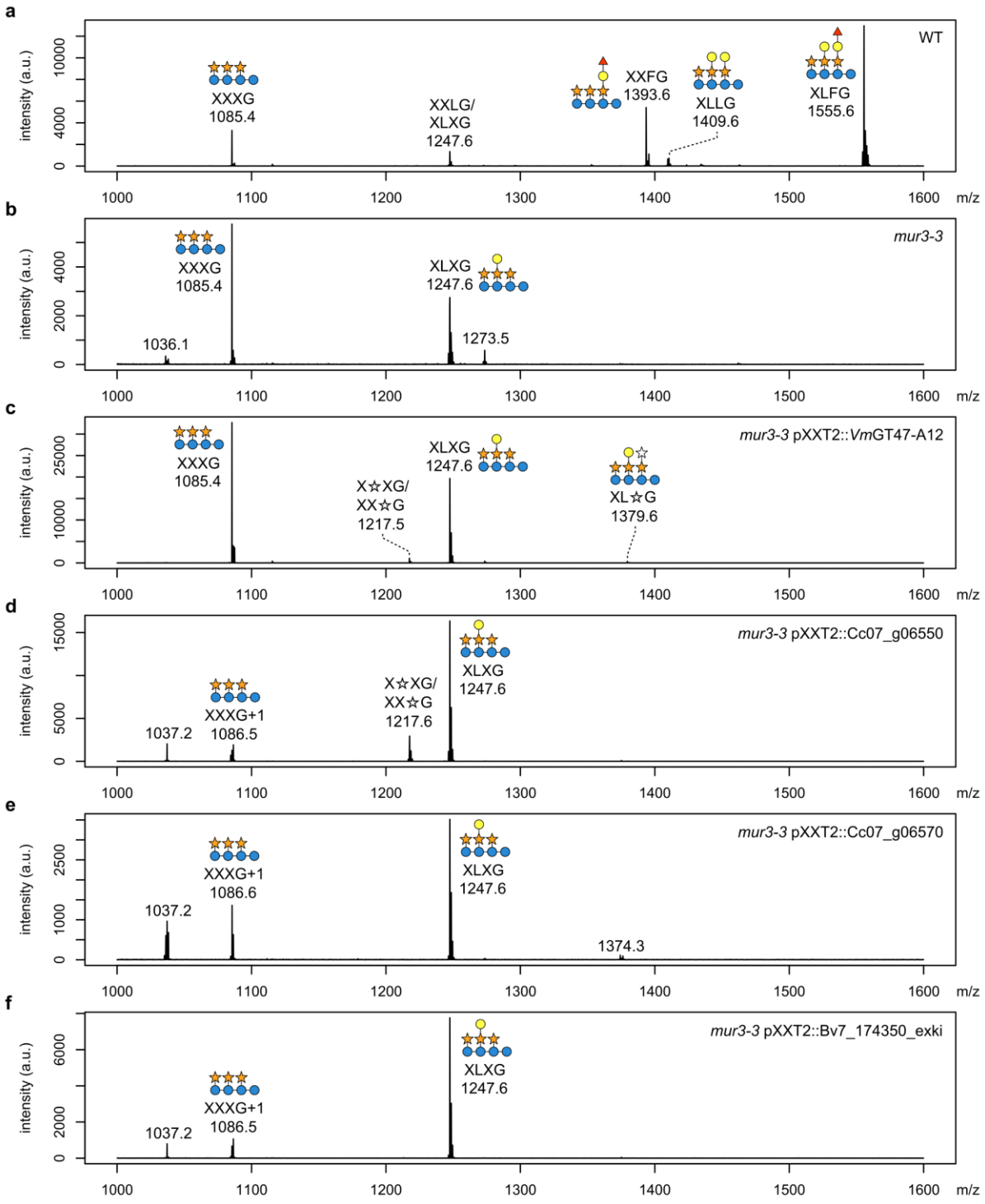


Figure 6.10 Expression of *Vm*GT47-A12 or Cc07_g06550 in *mur3-3* appears to produce a new pentosyl substituent on xyloglucan. Leaf AIR was treated with alkali to extract hemicellulose for subsequent digestion with the xyloglucanase XG5. The underivatised products were analysed by MALDI-TOF MS. Spectra correspond to **a** wild-type Arabidopsis, **b** *mur3-3* Arabidopsis, and *mur3-3* expressing **c** *Vm*GT47-A12, **d** Cc07_g06550, **e** Cc07_g06550, or **f** Bv7_174350_exki under the XXT2 promoter. For each construct, the results for only one independent line are shown (line #1 in each case); however, two further lines were also analysed per construct (see main text).

6.2.5 Over-expression of Cc07_g06550 and Vm*GT47-A12 in the* xlt2 mur3-1 *mutant dramatically alters xyloglucan structure*

I also analysed the xyloglucan structures of the *xlt2 mur3-1* transgenic lines by the same two methods. As before, PACE analysis demonstrated the presence of XXXG, XXLG, XXFG, XLLG, and XLFG in the wild type control (**Figure 6.11**). In contrast, XXXG was by far the most abundant product from the *xlt2 mur3-1* double mutant, with only a minor contingent of XXFG (estimated to constitute 4–6 % of the total subunits, $n = 3$). In all three *xlt2 mur3-1* lines over-expressing *Sl*XST1, however, the intensity of the XXXG band was dramatically reduced, with the appearance of a new, intense band at a somewhat higher position. A second, weak band was also visible below the position of XXFG. Taking into account the previously published activity of this enzyme (Schultink *et al,* 2013), these results suggest that almost all XXXG subunits in these plants were converted to arabinofuranose-containing XXSG subunits. Similarly, the XXXG band in the lines over-expressing Cc07_g06550 was also dramatically reduced, apparently to be replaced by another new, intense band, situated roughly at the position of XXLG. However, this band did not co-migrate with the band seen for the lines over-expressing *Sl*XST1. These results clearly demonstrate that over-expression of *Sl*XST1 or Cc07_g06550 radically changes the xyloglucan structure in *xlt2 mur3-1* plants, but also suggest that these two enzymes differ in their precise activity. In contrast, the xyloglucanase products from plants over-expressing *Vm*GT47-A12 were similar to those from untransformed *xlt2 mur3-1* plants. Nonetheless, a weak band was consistently visible for these lines, which co-migrated with XXSG. These results confirm that *Vm*GT47-A12[Δ167–193] does in fact have an activity on xyloglucan, albeit a weak one.

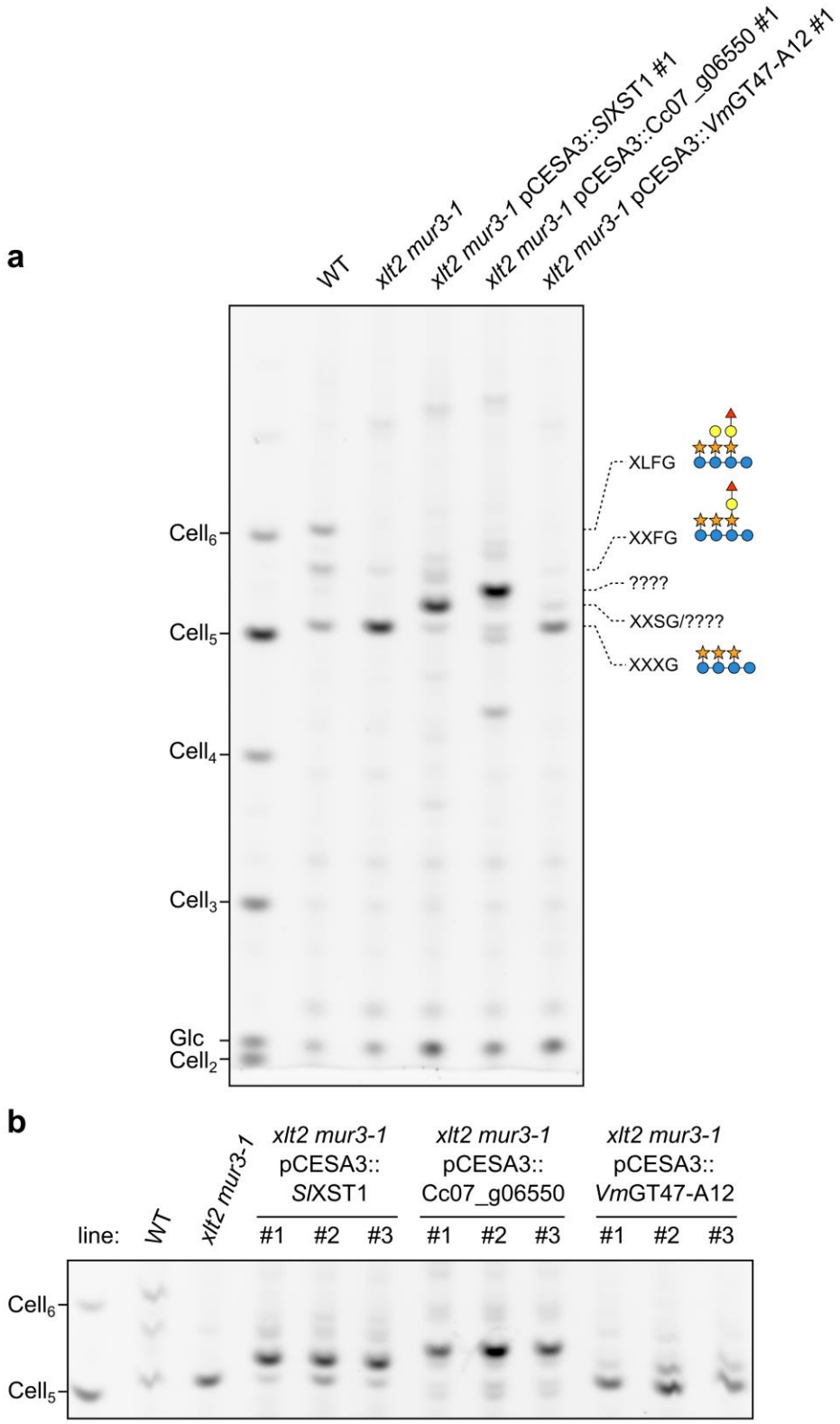


Figure 6.11 Overexpression of *Sl*XST1 or Cc07_g06550, but not *Vm*GT47-A12, in *xlt2 mur3-1* radically alters xyloglucan composition. Alkali-extracted hemicellulose was treated with XG5 xyloglucanase; the products were analysed by PACE. **a** PACE gel highlighting differences between wild-type Arabidopsis (WT), *xlt2 mur3-1*, and the transgenic lines (line #1 for each). **b** Confirmation of the same results in three independent lines per construct.

For MALDI-TOF MS analysis, only one line was analysed for each construct. As above, I prepared xyloglucanase products for mass spectrometry and Dr Li Yu operated the mass spectrometer. The wild-type control yielded the XXXG, XLXG/XXLG, XXFG, XLLG, and XLFG products as before, albeit with major M+1 peaks for XXXG, XXFG, and XLFG (**Figure 6.12**). The *xlt2 mur3-1* double mutant exhibited only a single glycan (M+1) peak of 1086.4 (sodiated XXXG), with no other subunits sufficiently abundant for detection. In stark contrast, but in accordance with the results from PACE analysis, xyloglucanase products from the plant over-expressing *Sl*XST1 exhibited only a minor XXXG peak (1085.3 m/z), instead displaying much more intense peaks for ions with m/z ratios of 1217.3 and 1349.3 (sodiated H_4P_4 and H_4P_5, respectively). These two ions likely correspond to the arabinofuranosylated oligosaccharides XXSG and XSSG, respectively. Since *Sl*XST1 is thought to transfer arabinofuranose only to the third xylose in the Arabidopsis XXXG repeat, these results may indicate that the xyloglucan in these plants is close to saturation in terms of its arabinofuranosylation. Similarly, the plant overexpressing Cc07_g06550 yielded the same 1217.3 m/z ion and exhibited only a weak 1085.3 m/z peak; however, the intensity of the peak at 1349.3 m/z was much lower than that for *Sl*XST1. These results agree with the previous MS experiment in suggesting that Cc07_g06550 is a xyloglucan pentosyltransferase; furthermore, they suggest that Cc07_g06550 may be more specific than *Sl*XST1 in terms of the position within the XXXG repeat at which it acts. In contrast, the plant over-expressing *Vm*GT47-A12 exhibited a mass profile similar to that of *xlt2 mur3-1*, except that low intensity peaks were visible for ions with m/z ratios of 1217.2 and 1394.3 in the *Vm*GT47-A12-expressing plant. While the latter ion is presumed to correspond to XXFG from the *xlt2 mur3-1* background, the former, as before, likely corresponds to X☆XG or XX☆G. These results support the notion that *Vm*GT47-A12[Δ167–193] exhibits a weak xyloglucan pentosyltransferase activity in Arabidopsis.

6.2.6 Xyloglucan structure can be analysed through the use of exo*-acting glycosidases*

Besides OLIMP and PACE, xyloglucan structures have also been analysed by HPAEC-PAD (Vincken *et al*, 1995; Konishi *et al*, 1998). However, the requirement of these analyses for pre-characterised standards impedes the characterisation of novel structures. On the other hand, although NMR can be used to definitively determine the structures of novel xyloglucan-derived oligosaccharides, it requires relatively large amounts of sample in high purity. Therefore, I sought an alternative method to characterise the products of Cc07_g06550 and *Vm*GT47-A12 activity further.

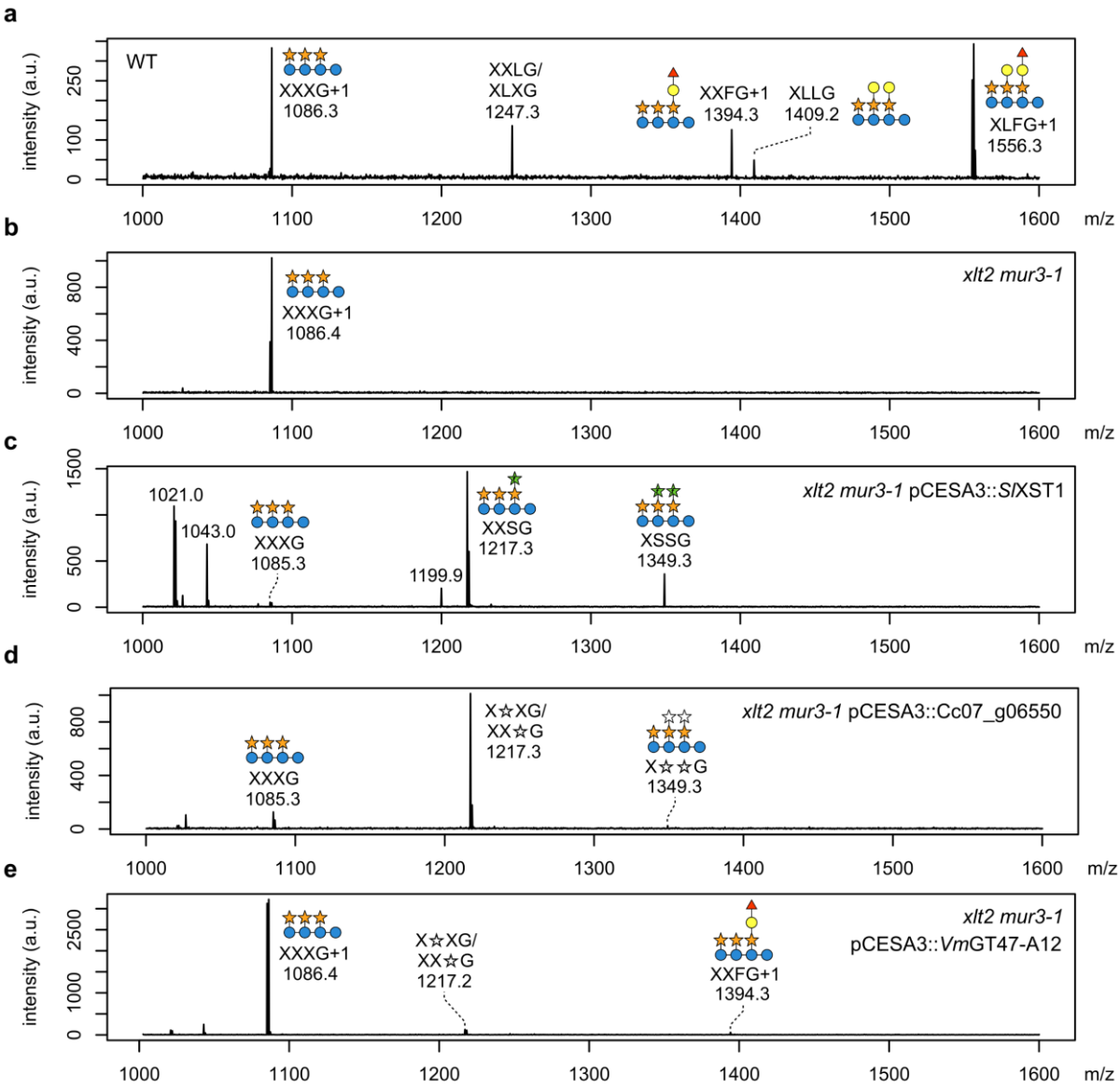


Figure 6.12 The products of Cc07_g06550 and *Vm*GT47-A12 activity when expressed in *xlt2 mur3-1* are similar to those produced in *mur3-3*. Leaf AIR was treated with alkali to extract hemicellulose for subsequent digestion with the xyloglucanase XG5. The underivatised products were analysed by MALDI-TOF MS. Spectra correspond to **a** wild-type Arabidopsis, **b** *xlt2 mur3-1* Arabidopsis, and *xlt2 mur3-1* expressing **c** *Sl*XST1, **d** Cc07_g06550, or **e** *Vm*GT47-A12 under the CESA3 promoter. Only one independent line was analysed per construct.

The structures of *endo*-glycosidase products from xylan and glucomannan, as well as other glycans, are now routinely analysed through the use of *exo*-glycosidases: enzymes that remove a single monosaccharide residue from a non-reducing terminus in their substrate (Kattla *et al*, 2011; Busse-Wicher *et al*, 2016; Tryfona *et al*, 2019; Yu *et al*, 2021a). For instance, in **Chapter 5**, I analysed GH11 and GH30 *endo*-xylanase products using *exo*-acting glucuronidases, galactosidases, and xylosidases—analysing the secondary products with PACE. However, although many bacterial xyloglucan-degrading enzymes have now been characterised (Larsbrink *et al*, 2011, 2014a, 2014b), analogous approaches do not yet appear widespread in analyses of xyloglucan structure. Therefore, I wanted to develop such a system as a means to analyse the products of xyloglucan-specific GT47-As.

Wild-type Arabidopsis xyloglucan contains β1,4-linked glucosyl, α1,6-linked xylosyl, β1,2-linked galactosyl, and α1,2-linked fucosyl residues; hence, I tested the abilities of β-glucosidase, α-xylosidase, β-galactosidase, and α-fucosidase enzymes to disassemble xyloglucanase products into their constituent monosaccharides. First, I extracted hemicellulose from Arabidopsis leaf AIR using alkali and digested the preparation with GH5 xyloglucanase. After removal of enzyme and residual polysaccharides by precipitation in 65 % ethanol, the resultant oligosaccharides were subject to digestion by a combination of *Meripilus* sp. β-galactosidase (GH family unknown), *Bifidobacterium bifidum* AfcA GH95 α1,2-fucosidase, *E. coli* GH31 α-xylosidase, and/or *Aspergillus niger* GH3 β-glucosidase. Enzymes were removed by 65 % ethanol precipitation before the addition of β-glucosidase. As above, ANTS-derivatised products were then analysed by PACE analysis. The results provided clear evidence for the activity for all four enzymes (**Figure 6.13**). Once again, xyloglucanase digestion of wild-type xyloglucan produced the XXXG, XXLG, XLXG, XXFG, XLLG, and XLFG subunits. However, further digestion of these products with β-galactosidase resulted in the complete loss of the XXLG, XLXG, XLLG, and XLFG bands, with an increase in the abundance of XXXG and XXFG; analogously, digestion with α-fucosidase resulted in the complete loss of XXFG and XLFG, and an increase in the abundance of XXLG and XLLG. As anticipated, co-digestion with α-fucosidase and β-galactosidase resulted in the conversion of all higher xyloglucanase products to XXXG. Inclusion of β-glucosidase did not result in any additional change in the absence of α-xylosidase, confirming that glucosyl residues in the XXXG oligosaccharide are not accessible to β-glucosidase activity. In contrast, when α-xylosidase was included, the XXXG oligosaccharide shifted downwards; mass spectrometry confirmed that this was due to the loss of a single pentose (data not shown). Since other GH31

α-xylosidases have been shown to remove only the xylose from the glucose at the non-reducing end of XXXG (Fanutti *et al*, 1991; Moracci *et al*, 2000; Günl & Pauly, 2011; Larsbrink *et al*, 2011, 2014a), this product very likely constitutes GXXG. Indeed, the structure of this oligosaccharide was confirmed by the fact that, in contrast to XXXG, the GXXG oligosaccharide lost a single hexosyl residue after digestion with β-glucosidase (confirmed by mass spectrometry; data not shown). Further sequential digestion steps resulted in the complete

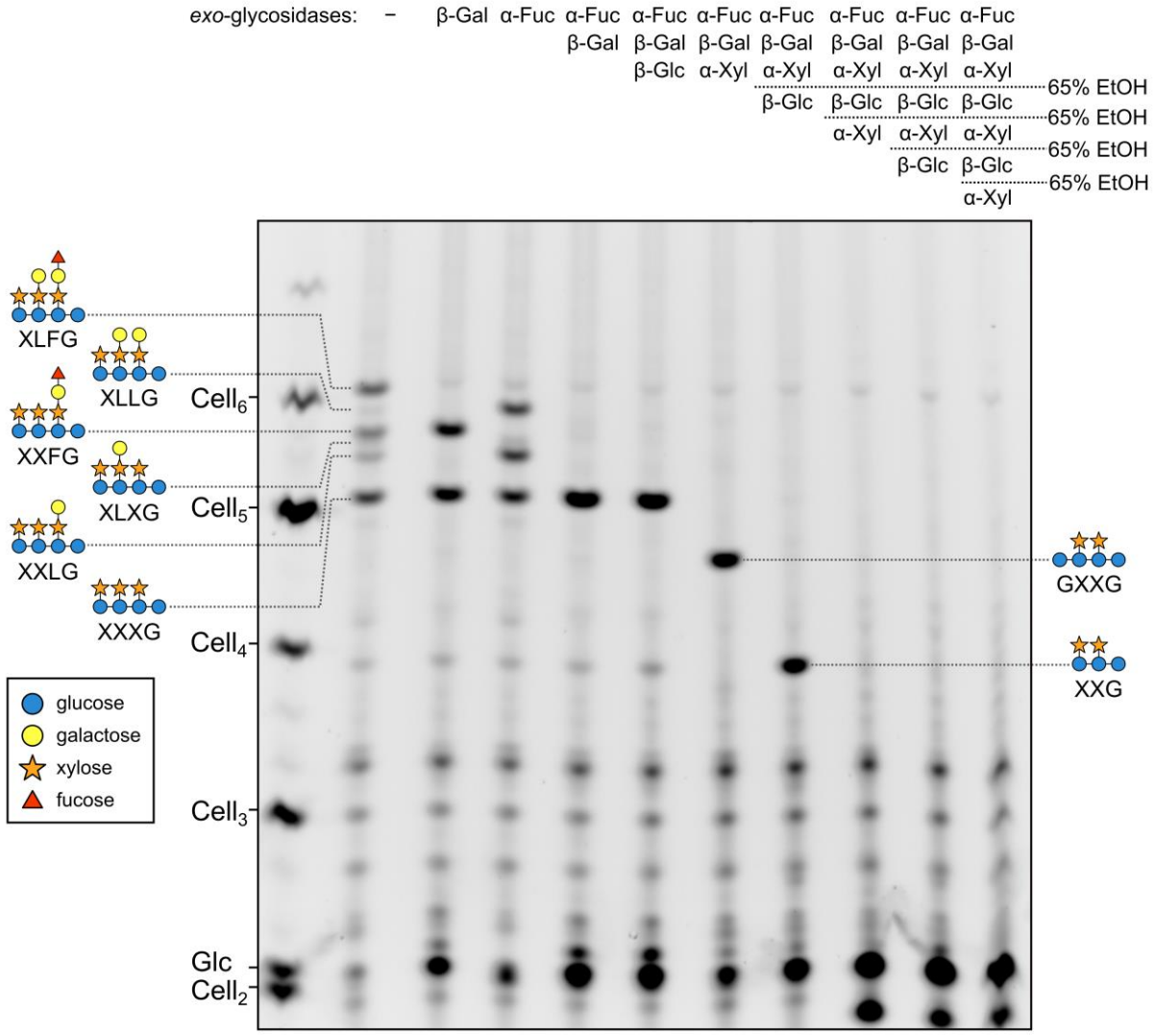


Figure 6.13 Arabidopsis xyloglucan can be degraded sequentially be treatment with *exo*-glycosidases. Leaf AIR was treated with alkali to extract hemicellulose for subsequent digestion with the xyloglucanase XG5. After removal of xyloglucanase with 65 % ethanol precipitation, samples were treated with a combination of *exo*-glycosidases before analysis by PACE. β-Gal: β-galactosidase from *Meripilus* sp.; α-Fuc: *Bb*AfcA α-fucosidase; β-Glc: *An*GH3 β-glucosidase; α-Xyl: *Ec*GH31 α-xylosidase. The dotted line represents removal of enzymes by precipitation in 65 % ethanol.

conversion of all oligosaccharides to their four monosaccharide constituents, presumably due to the failure of 65 % ethanol precipitation to adequately inactivate or remove β-glucosidase.

6.2.7 Cellvibrio japonicus *GH51 α-arabinofuranosidase and* Chætomium globosum *GH3 β1,2-xylosidase exhibit activity on xyloglucan from tomato and blueberry, respectively*

I also wanted to identify further *exo*-glycosidases that could be used to distinguish different pentosyl-xylose sidechains, such as the S, U, and D sidechains. However, since Arabidopsis xyloglucan does not contain these structures, I needed to identify some alternative substrates for activity screens. *Solanum lycopersicum* (tomato; Solanales) and *V. corymbosum* (blueberry; Ericales) fruits posed useful sources of xyloglucan decorated with α-arabinofuranose and β-xylose, respectively. Accordingly, I digested alkali-extracted hemicellulose from tomato and blueberry fruit AIR with GH5 xyloglucanase to release the xyloglucan subunits. PACE analysis of the ANTS-derivatised oligosaccharides revealed that the products differed substantially to those released from Arabidopsis xyloglucan (**Figure 6.14**). A wide variety of products were released from tomato xyloglucan; however, no band co-migrating with the Arabidopsis XXXG band was present. These results were anticipated, as tomato xyloglucan has been shown to be possess a complex structure based primarily on an XXGG repeat (Jia *et al*, 2003; Hoffman *et al*, 2005). In contrast, the blueberry xyloglucanase products constituted only five major bands, with the smallest co-migrating with the Arabidopsis XXXG (confirmed in multiple gels—data not shown), and the third highest co-migrating with XXFG, suggesting that blueberry xyloglucan is built on an XXXG repeat—just as it is in the closely related *V. myrtillus* L. (bilberry) (Hilz *et al*, 2007).

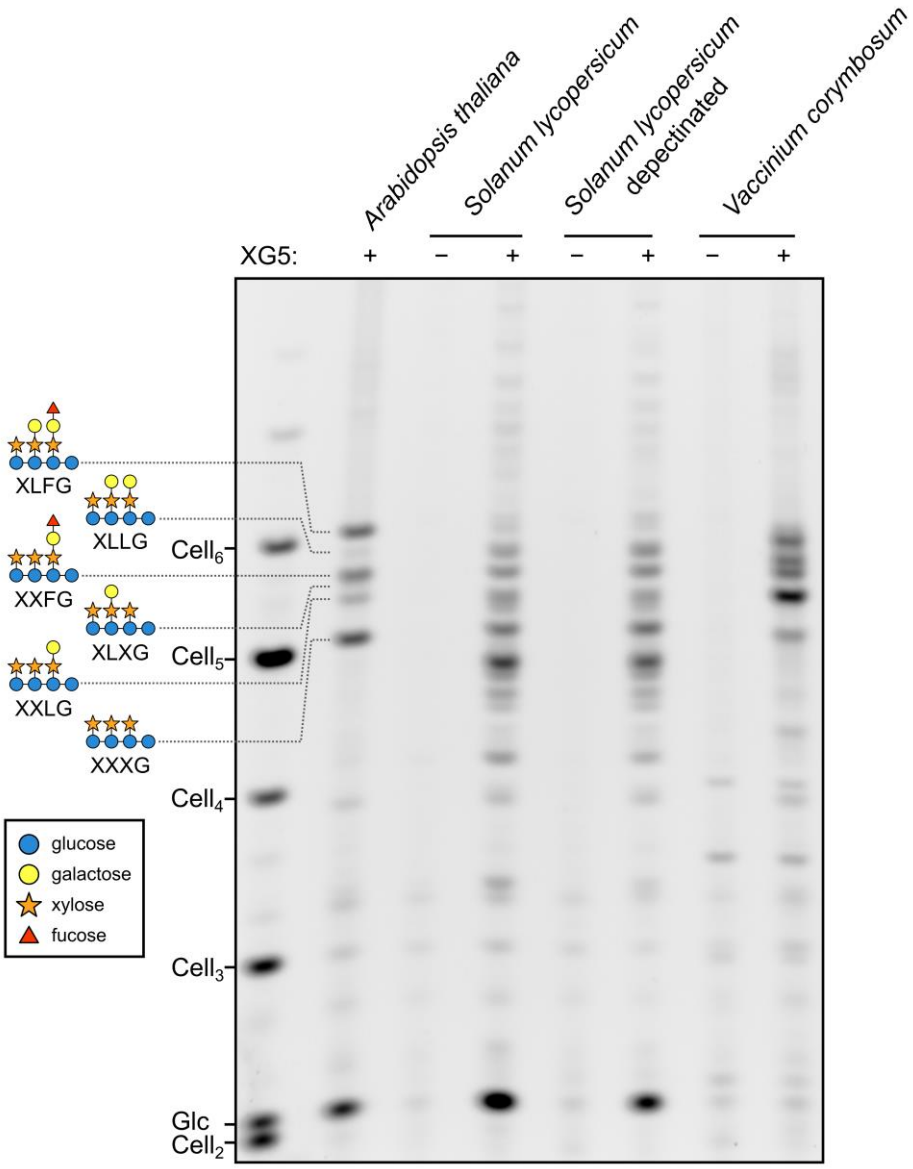


Figure 6.14 Xyloglucan subunits from *Solanum lycopersicum* and *Vaccinium corymbosum* fruits exhibit different structures to those from Arabidopsis leaf. Hemicellulose was extracted from Arabidopsis leaf AIR, *Solanum lycopersicum* fruit AIR, ammonium oxalate-treated *S. lycopersicum* fruit AIR, and *Vaccinium corymbosum* fruit skin AIR using alkali before digestion with XG5 xyloglucanase. Products were derivatised with ANTS and analysed by PACE.

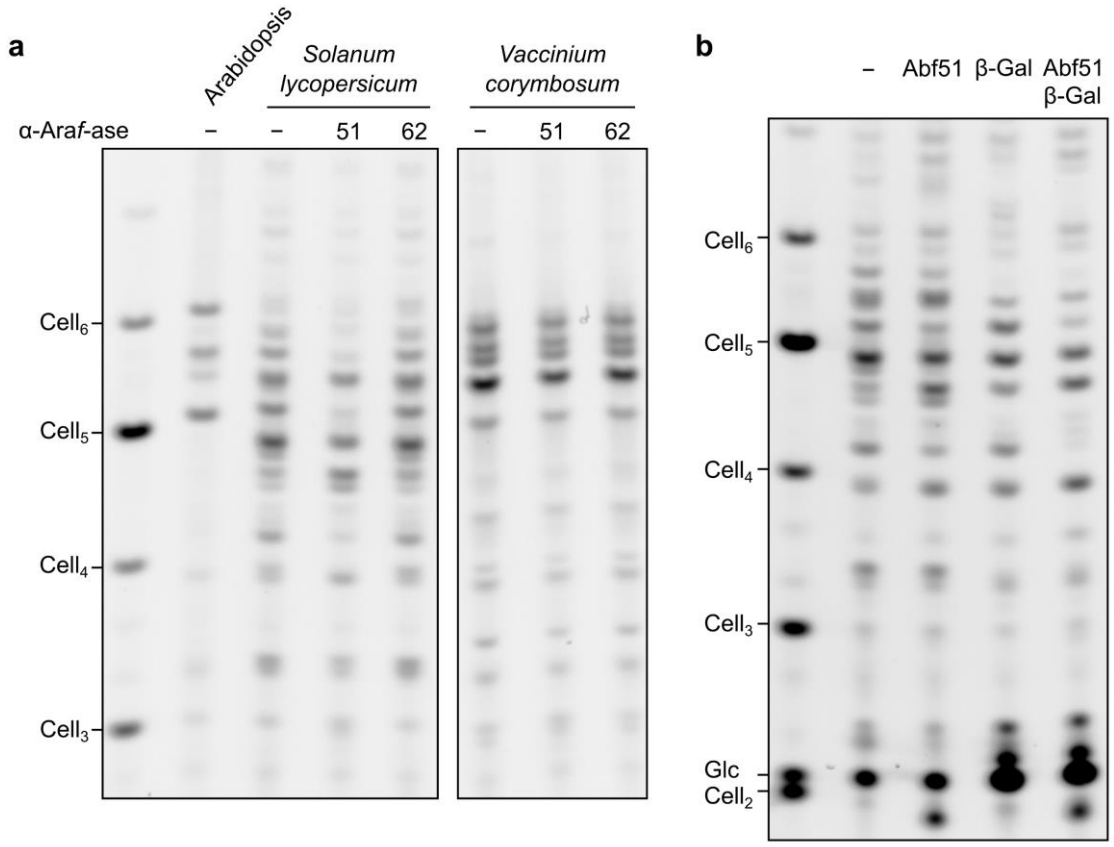


Figure 6.15 *Cj*Abf51 α-arabinofuranosidase exhibits weak activity on xyloglucan from *Solanum lycopersicum*, but not that from Arabidopsis or *Vaccinium corymbosum*. Alkali-extracted hemicellulose from Arabidopsis leaf AIR, *Solanum lycopersicum* fruit AIR, and *Vaccinium corymbosum* fruit skin AIR was digested with XG5 xyloglucanase. Xyloglucanase was subsequently removed by precipitation in 65 % ethanol. **a** The xyloglucanase products were incubated overnight with no enzyme, *Cj*Abf51 GH51 α-arabinofuranosidase ('51'), or *Pa*GH62 α-arabinofuranosidase/β-xylosidase ('62') before PACE analysis. **b** *S. lycopersicum* xyloglucanase products were incubated overnight with a combination of *Cj*Abf51 GH51 α-arabinofuranosidase and/or β-galactosidase from *Meripilus* sp. ('β-Gal') before PACE analysis.

To identify an α-arabinofuranosidase capable of acting specifically on the S sidechains of tomato xyloglucan, I carried out a secondary digestion on these xyloglucanase products using two candidate α-arabinofuranosidases: Abf51 GH51 α-arabinofuranosidase from *Cellvibrio japonicus* and GH62 xylan α-arabinofuranosidase/β1,2-xylosidase from *Penicillum aurantiogriseum*. PACE analysis revealed that digestion with *Cj*Abf51 resulted in a substantial change in the ratios of the different xyloglucanase products from tomato fruit (though the identity of these bands could not be assigned; **Figure 6.15a**). In contrast, the GH62 enzyme had no activity on these oligosaccharides. Neither enzyme had any substantial activity on the

blueberry xyloglucanase products. Subsequently, I attempted to characterise the tomato oligosaccharides better by digesting similar xyloglucanase products with a combination of *Cj*Abf51 and/or β-galactosidase. The results of the PACE analysis of the products were consistent with the idea that the two enzymes work on different substrates, as different bands were affected by each enzyme (**Figure 6.15b**). However, *Cj*Abf51 activity was inconsistent in its magnitude across experiments, and the enzyme appeared unable to completely eliminate its apparent substrates (side activities in the enzyme preparation prevented longer incubations times or larger enzyme:substrate ratios). These results suggest that *Cj*Abf51 has weak α1,2-arabinofuranosidase activity on xyloglucan.

I also tested the activity of a GH3 β1,2-xylosidase from *Chætomium globosum* (*Cg*GH3*)* on the xyloglucanase products from blueberry. This enzyme has previously been used to remove β1,2-linked xylose from the $D^{2,3}$ structure found in grass xylan (Xyl-β1,2-Ara-α1,3-Xyl) (Tryfona *et al*, 2019). Overnight incubation with this enzyme resulted in a subtle shift in the ratios between the different blueberry xyloglucanase products, and a much longer incubation time allowed the complete removal of the two highest main bands visible by PACE (**Figure 6.16a**). These results suggest that this xylosidase may also indeed exhibit xylosidase activity against blueberry xyloglucan, but only at a very low level. Therefore, I increased the ratio of enzyme:substrate in subsequent reactions. However, further experiments revealed that this enzyme preparation also exhibited a low level of β-galactosidase activity (data not shown); hence, subsequent results were analysed carefully, with this fact in mind.

To better characterise these β-xylosidase-sensitive oligosaccharides, I digested similar blueberry xyloglucanase products with a combination of *Cj*Abf51 α-arabinofuranosidase, *Cg*GH3 β1,2-xylosidase, α1,2-fucosidase, β-galactosidase, and/or α-xylosidase. PACE analysis of the ANTS-derivatised products revealed that as before, *Cg*GH3 β1,2-xylosidase was able to remove the two highest of the five major bands (‘band 4’ and ‘band 5’), with a concomitant increase in intensity of the lowest and third highest bands (‘band 1’ and ‘band 3’), thought to constitute XXXG and XXFG, respectively (**Figure 6.16b**). Band 3 and 5 were both were sensitive to α1,2-fucosidase, whereas the intensity of band 4 was increased. Put together, these results suggest that band 5 might constitute XUFG, whereas band 4 might constitute any or a combination of XULG, XUUG, and XLUG. Combining both β1,2-xylosidase and α1,2-fucosidase with β-galactosidase resulted in the collapse of all bands to bands 1 and 2; none of the bands, before or after this digestion, were sensitive to *Cj*Abf51. Therefore, the identity of band 2 remained to be determined.

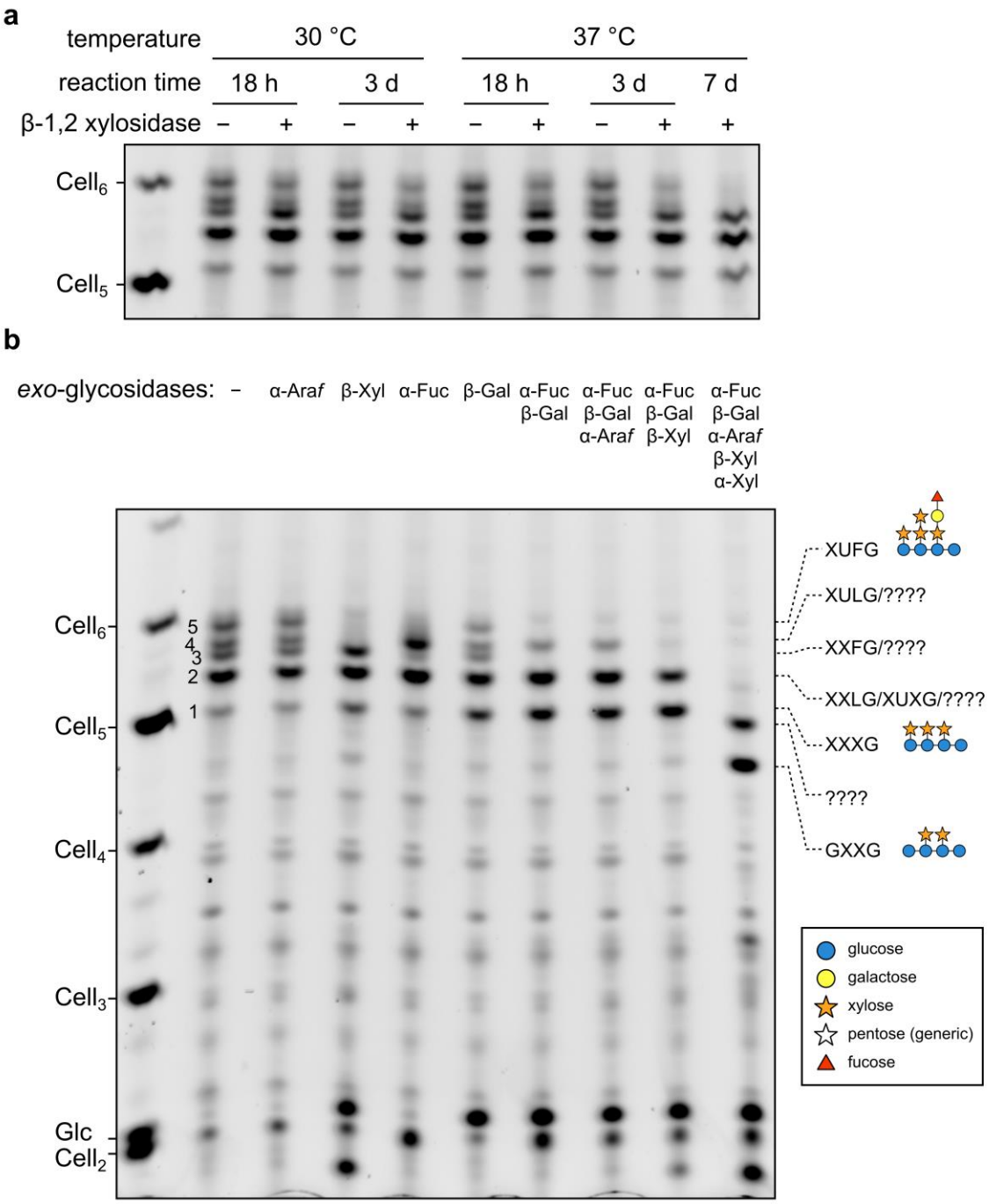


Figure 6.16 *Cg*GH3 β1,2-xylosidase can be used to probe the structure of xyloglucan from *Vaccinium corymbosum*. Alkali-extracted hemicellulose from *Vaccinium corymbosum* fruit skin AIR was digested with XG5 xyloglucanase. Xyloglucanase was subsequently removed by precipitation in 65 % ethanol. **a** Xyloglucanase products were incubated with either no enzyme or with *Cg*GH3 β1,2-xylosidase for a range of time periods at either 30 or 37 °C. **b** Xyloglucanase products were treated with a combination of *exo*-glycosidases. α-Ara*f*: *Cj*Abf51 α-arabinofuranosidase; β-Xyl: *Cg*GH3 β1,2-xylosidase (higher concentration and long incubation), α-Fuc: *Bb*AfcA α1,2-fucosidase, β-Gal: β-galactosidase from *Meripilus* sp.; α-Xyl: *Ec*GH31 α-xylosidase. Both experiments were analysed by PACE.

Consequently, a subset of these reactions was also analysed by MALDI-TOF MS. As before, I prepared xyloglucanase products for mass spectrometry and Dr Li Yu operated the mass spectrometer. Without the addition of any *exo*-glycosidases, the xyloglucanase products were revealed to be more complex in their make-up than had been revealed by PACE (**Figure 6.17**). Nevertheless, the main five ions detected (1084.8, 1216.9, 1247.0, 1379.0, and 1525.1 m/z) could plausibly be assigned to the five PACE bands—based on their masses, these ions likely represent sodium adducts of XXXG, X☆XG/XX☆G, XXLG/XLXG, X☆LG/ XL☆G, and XUFG, where, as before, '☆' represents a glucose decorated with a pentosyl-xylose disaccharide sidechain such as U, S, or D. Furthermore, after treatment with *Cg*GH3 β1,2-

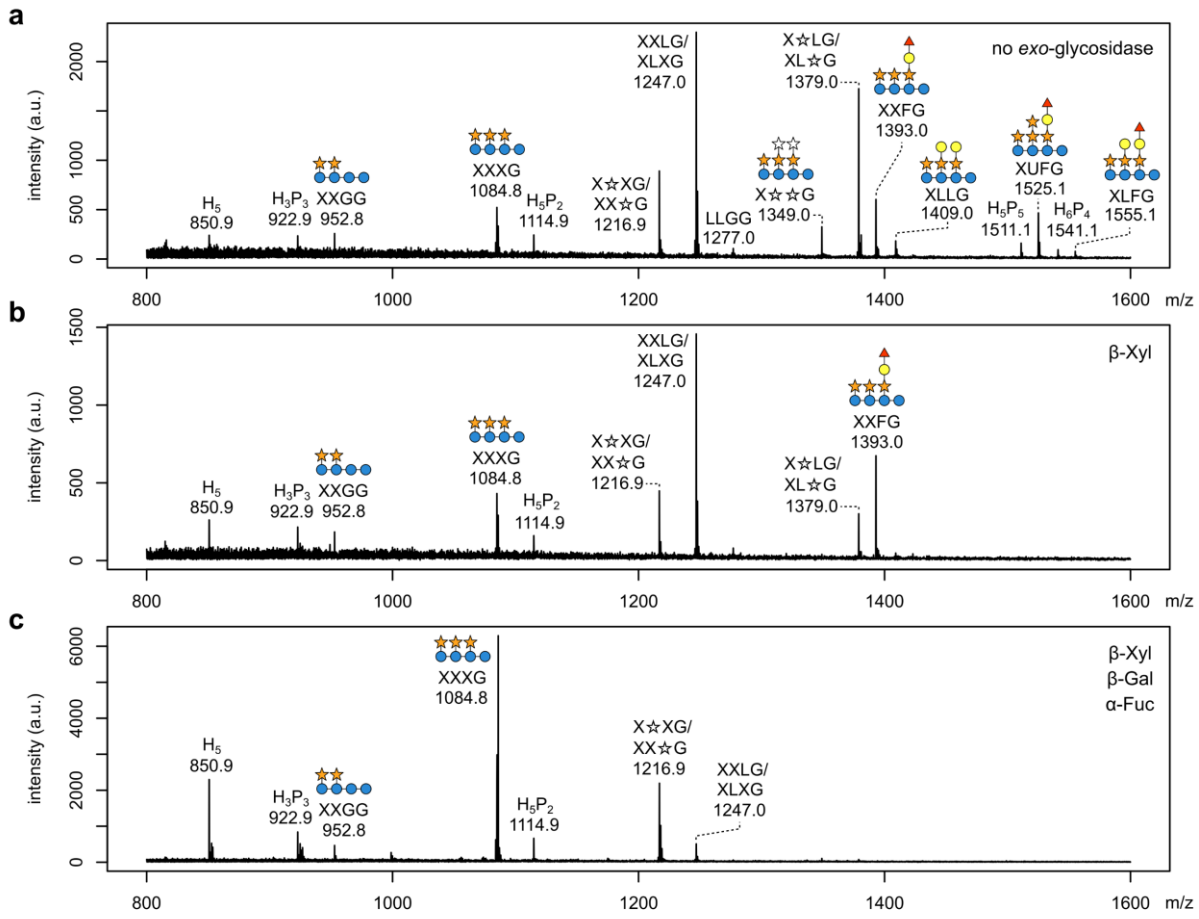


Figure 6.17 *Cg*GH3 β1,2-xylosidase appears to remove pentose sidechains from *Vaccinium corymbosum* xyloglucan. Alkali-extracted hemicellulose from *Vaccinium corymbosum* fruit skin AIR was digested with XG5 xyloglucanase. Xyloglucanase was subsequently removed by precipitation in 65 % ethanol. Xyloglucanase products were then treated with a combination of *exo*-glycosidases. β-Xyl: *Cg*GH3 β1,2-xylosidase (higher concentration and long incubation), α-Fuc: *Bb*AfcA α1,2-fucosidase, β-Gal: β-galactosidase from *Meripilus* sp.. The products were analysed by MALDI-TOF MS.

xylosidase, the 1349.0 and 1525.0 m/z ions, which likely correspond to X☆☆G and XUFG, completely disappeared, confirming that the effects of this enzyme are not limited to the β-galactosidase side activity. However, while reduced somewhat in intensity, neither the 1216.9 m/z peak (sodiated H_4P_4) nor the 1379.0 m/z peak (sodiated H_5P_3) was removed by β-xylosidase treatment. Furthermore, after combined treatment with β1,2-xylosidase, α1,2-fucosidase, and β-galactosidase, the 1216.9 m/z peak was still not digested; hence, this ion most likely corresponds to band 2 in the PACE gel. These data suggest that either the *Cg*GH3 β1,2-xylosidase is not able to access β-xylosyl residues at some positions in the XXXG repeat, or that blueberry xyloglucan possesses a non-xylose, hitherto-undocumented pentosyl substituent, such as arabinopyranose.

6.2.8 Cc07_g06550 likely encodes a xyloglucan β-xylosyltransferase whereas Vm*GT47-A12 likely encodes a xyloglucan α-arabinofuranosyltransferase*

Having identified a putative xyloglucan α-arabinofuranosidase and a putative xyloglucan β-xylosidase, I attempted to use these enzymes to determine the donor sugar specificities of Cc07_g06550 and *Vm*GT47-A12. Accordingly, I treated the GH5 xyloglucanase products from the transgenic *xlt2 mur3-1* plants with these enzymes and analysed the products using PACE. For the *Sl*XST1- and *Vm*GT47-A12-expressing plants, treatment with *Cj*Abf51 α-arabinofuranosidase resulted in either a reduction in intensity or the complete disappearance of XXSG / the novel band migrating above XXXG, with a concomitant increase in the intensity of the band corresponding to XXXG (**Figure 6.18**). In contrast, treatment with *Cg*GH3 β-xylosidase had extremely little, if any, effect. The same results were replicated when I repeated the experiment with a range of *Cj*Abf51 concentrations (data not shown). Conversely, treatment with α-arabinofuranosidase had no effect on the novel band seen for the Cc07_g06550-expressing plant, whereas the same band appeared to be almost completely converted to XXXG by β-xylosidase treatment. These results suggest that these two different bands could contain terminal α-arabinofuranose and β-xylose respectively, and therefore that Cc07_g06550 and *Vm*GT47-A12 could encode a β-xylosyltransferase and an α-arabinofuranosyltransferase, respectively.

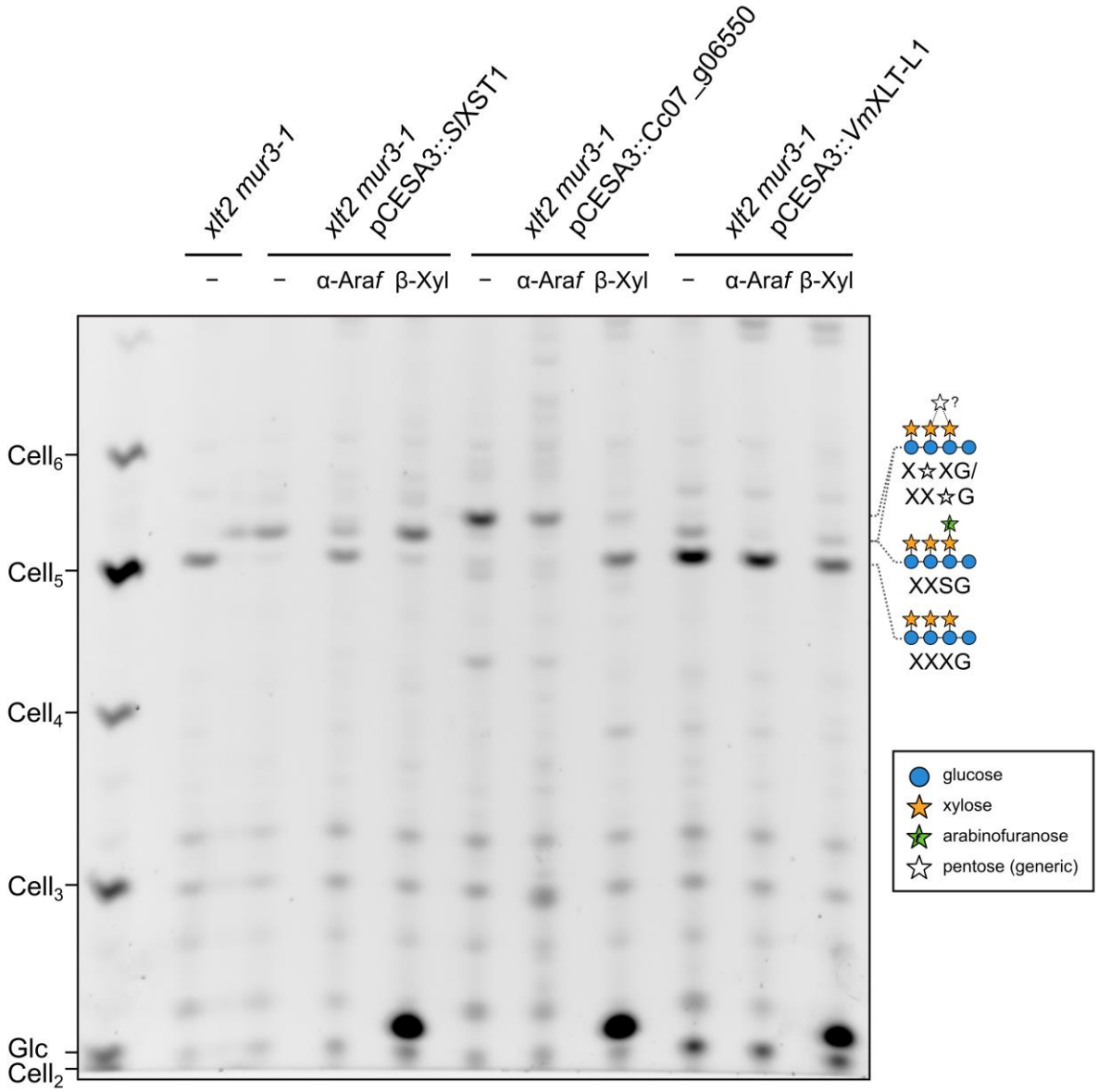


Figure 6.18 The products of *Sl*XST1 and *Vm*GT47-A12 are sensitive to α-arabinofuranosidase, whereas the product of Cc07_g06550 is sensitive to β-xylosidase. Alkali-extracted hemicellulose from leaf AIR of *xlt2 mur3-1* transgenic lines was digested with XG5 xyloglucanase. Xyloglucanase was subsequently removed by precipitation in 65 % ethanol. The products were then treated with either no enzyme, *Cj*Abf51 α-arabinofuranosidase ('α-Ara*f*'), or *Cg*GH3 β-xylosidase ('β-Xyl'). The products were derivatised with ANTS and analysed by PACE. Only one independent line was analysed for each construct.

6.2.9 Cc07_g06550 transfers a sugar to the second xylose in XXXG whereas Vm*GT47-A12 transfers a sugar to the third*

The large difference in migration between the two different novel bands suggested that the positioning of the novel decoration might vary between them. In fact, I previously developed a method to determine the position of secondary substitutions within the XXXG repeat. This method relies on the fact that α-xylosidase-treated xyloglucanase products (which have an unsubstituted glucosyl residue at the non-reducing terminus) are protected from *An*GH3 β-glucosidase if the glucose adjacent to the non-reducing end is substituted with a disaccharide sidechain (*i.e.* GXLG is sensitive to β-glucosidase but GLXG is not; data not shown). To put this method into action, I treated xyloglucanase products from untransformed and transgenic *xlt2 mur3-1* plants with α-xylosidase and β-glucosidase, and analysed the ANTS-derivatised products using PACE. As before, products from the plants over-expressing *Sl*XST1, Cc07_g06550, or *Vm*GT47-A12 were seen to exhibit two bands in the $Cell_5$–$Cell_6$ region, corresponding to XXXG and one larger subunit. Further inspection of the PACE gel revealed that, as expected, all of these oligosaccharides were sensitive to α-xylosidase (**Figure 6.19**). However, whereas for *Sl*XST1 and *Vm*GT47-A12, both α-xylosidase products were sensitive to β-glucosidase, the larger α-xylosidase product for Cc07_g06550 was not, indicating that the structure of this oligosaccharide impedes the action of the β-glucosidase enzyme. These results suggest that the pentosyl-xylose sidechain created by Cc07_g06550 is located at the second position in the XXXG repeat and is thus able to block the action of β-glucosidase. Furthermore, these results suggest that, in contrast, the pentosyl-xylose sidechains created by *Sl*XST1 and *Vm*GT47-A12 are likely to be located at the third position in the XXXG repeat, as they did not inhibit the removal of glucose from the non-reducing end of the oligosaccharide.

6.2.10 Pentosyl-xylose disaccharide decorations are difficult to identify in native xyloglucans

To date, xylosyl-xylose sidechains have only been detected in xyloglucans from plants in the Ericales order (Pauly & Keegstra, 2016; Zavyalov *et al*, 2019). This order constitutes an early-diverging group of the asterids—a major clade of eudicot plants (**Figure 1.15**). *C. canephora*, however, does not belong to the Ericales, instead constituting a member of the Gentianales order. Along with the Solanales, all three of these orders belong to the asterids. Hence, if Cc07_g06550 does indeed encode a xyloglucan β-xylosyltransferase, xylosyl-xylose sidechains (i.e. 'U' sidechains) could in fact be more widespread amongst asterids. More generally, since I found that most asterid genomes contain an XST homologue, pentosyl-xylose disaccharide sidechains ought to be detectable in most asterid species. Furthermore, close

inspection of my GT47-A tree revealed that a limited number of rosid species also appear to have XST homologues, suggesting that other rosids (such as Arabidopsis) have since lost this enzyme. Hence, these sidechains could also be present in the wider eudicots.

In order to investigate the prevalence of pentosyl-xylose xyloglucan sidechains across the eudicots, I prepared AIR from the fruit skin of *A. chinensis* (kiwi; Ericales), the leaves of the asterid *Artemisia dracunculus* (tarragon; Asterales), and the fruit skin of the rosid *Cucumis sativus* (cucumber; Cucurbitales). I also wanted to analyse the xyloglucan from *C. canephora*;

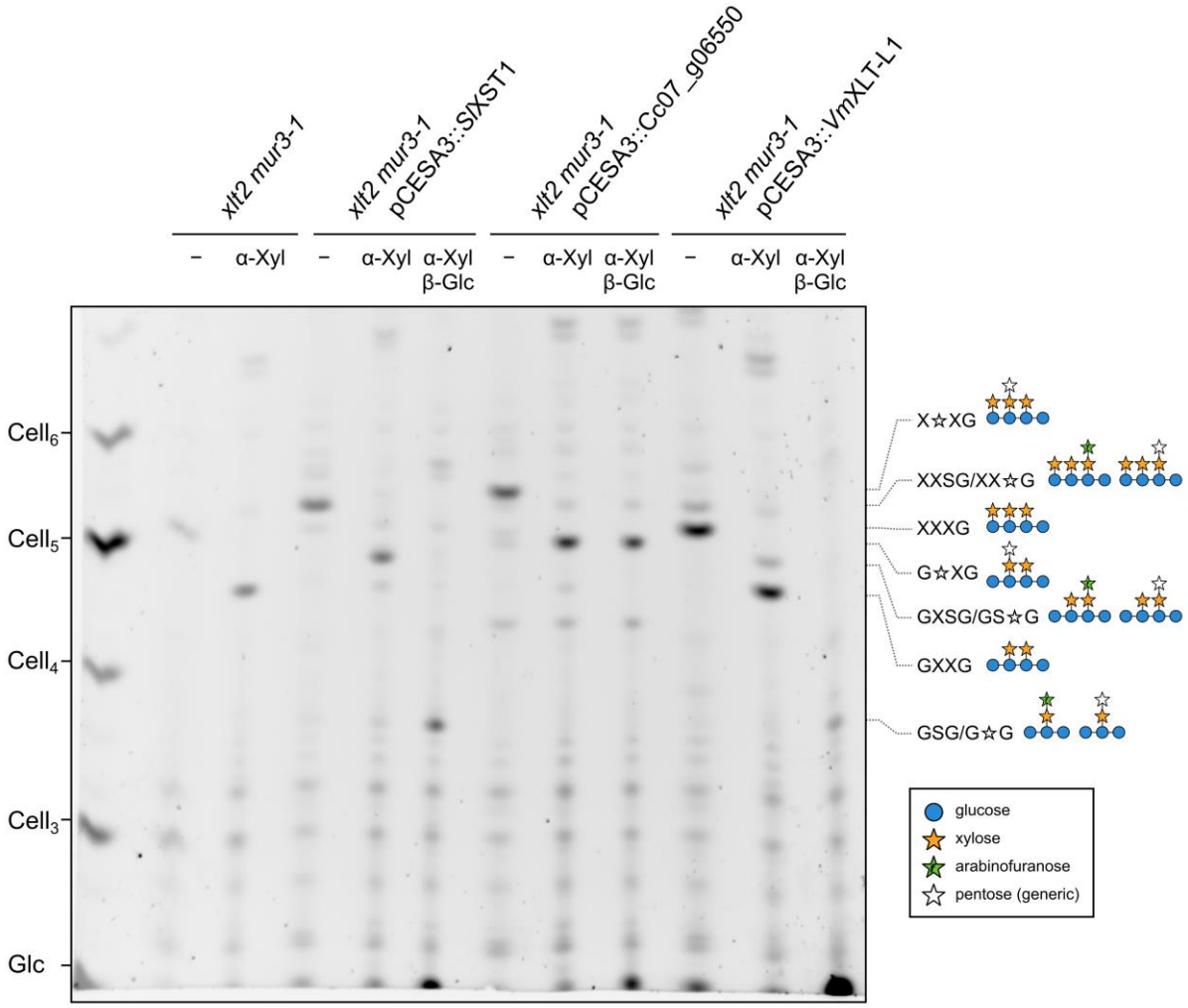


Figure 6.19 α-Xylosidase-treated xyloglucanase products from plants expressing *Sl*XST1 or *Vm*GT47-A12, but not Cc07_g06650, are sensitive to β-glucosidase. Alkali-extracted hemicellulose from leaf AIR of *xlt2 mur3-1* transgenic lines was digested with XG5 xyloglucanase. Xyloglucanase was subsequently removed by precipitation in 65 % ethanol. The products were then treated with either no enzyme, *Ec*GH31 α-xylosidase ('α-Xyl'), or α-xylosidase and *An*GH3 β-glucosidase ('β-Glc'). The products were derivatised with ANTS and analysed by PACE. Only one independent line was analysed for each construct.

however, due to lack of material, I prepared AIR from the leaves of close relative *Coffea arabica* 'Catimor' instead. As previously, I extracted hemicellulose from these preparations using alkali; this hemicellulose was then digested with GH5 xyloglucanase. The products were analysed by PACE alongside the previously generated xyloglucanase products from wild-type Arabidopsis, *xlt2 mur3-1* pCESA3:: Cc07_g06550, *xlt2 mur3-1* pCESA3::*Sl*XST1, *V. corymbosum*, and *S. lycopersicum*. Judging from the results, xyloglucan from *A. dracunculus* leaf and *C. sativus* fruit skin exhibited a similar structure to that of wild-type Arabidopsis, as strong bands co-migrating with XXXG, XXFG, and XLFG were visible (**Figure 6.20**). For both of these samples, two faint bands were also visible at and just below the position of XXLG; however, it was not clear whether these bands were the result of the xylanase side activities of this xyloglucanase. These results indicate that if pentosyl-xylose disaccharide sidechains are at all present in the xyloglucan in these plants, they are likely of very low abundance in the tissues analysed. In contrast, the subunits released from *A. chinensis* fruit skin xyloglucan appeared entirely different in character. The bands observed did not appear to co-migrate with any of those in the other samples, including the XXXG band. This, combined with the fact that many of the products migrated more quickly than XXXG, suggests that this xyloglucan does not exhibit a XXXG-based repeat. The pattern of bands somewhat resembled that of the *S. lycopersicum* xyloglucanase products in their spacing, however, the bands themselves did not co-migrate. Similarly, the products released from the *C. arabica* xyloglucan were difficult to assign. However, the amount of product released from this sample by GH5 xyloglucanase was very low, suggesting that xyloglucan may only make up a minor proportion of the cell wall in leaves from this plant. In a second experiment, the xyloglucanase concentration was increased tenfold; however, this did not result in increased amounts of xyloglucan-derived product (data not shown). Hence, it was not possible to determine whether the *A. chinensis* or *C. arabica* xyloglucans contained pentosyl-xylose disaccharide sidechains from these results. Due to time restrictions, it was not possible to investigate this further using the β-xylosidase or α-arabinofuranosidase enzymes.

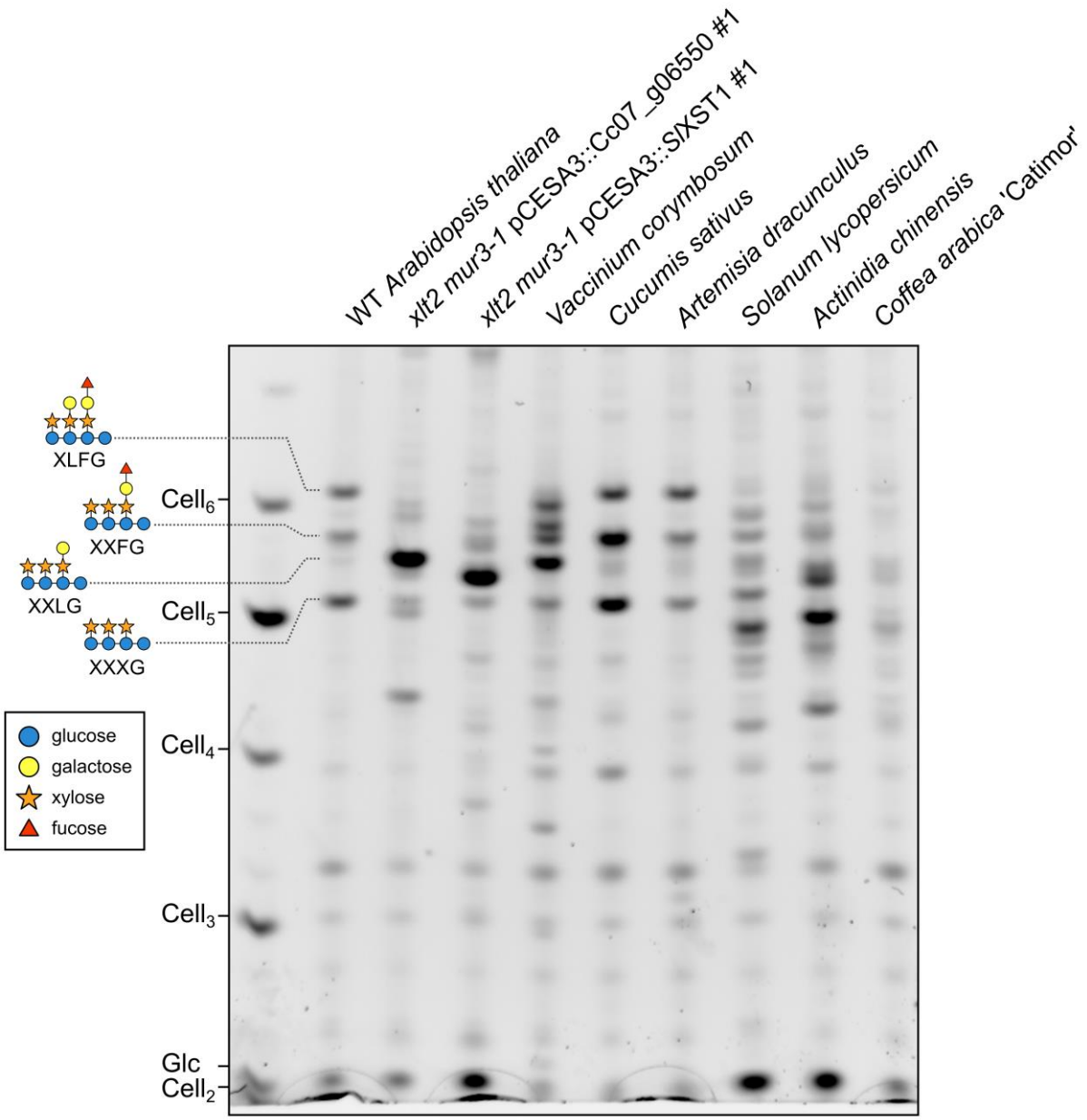


Figure 6.20 Variation in xyloglucan subunits between different eudicot plants. Hemicellulose was extracted from wild-type or transgenic *xlt2 mur3-1* Arabidopsis leaf AIR, *Vaccinium corymbosum* fruit skin AIR, *Cucumis sativus* fruit skin AIR, *Artemisia dracunculus* leaf AIR, *Solanum lycopersicum* fruit AIR, *Actinidia chinensis* fruit skin AIR, and *Coffea arabica* 'Catimor' leaf AIR using alkali. Extracted hemicellulose was digested with XG5 xyloglucanase, and the products were derivatised with ANTS and analysed by PACE.

6.3 Discussion

In this chapter, I have used MALDI-TOF MS and PACE combined with *exo*-glycosidase digestions to analyse xyloglucan structure. The latter appears to represent a somewhat novel approach, and exhibits several strengths that may be of use to future investigators. For a start, the sensitivity of PACE is ostensibly superior to that of MALDI-TOF MS, and is also at least comparable with that of HPAEC-PAD. For example, using PACE, I was able to detect that around 5 % of the xyloglucan subunits from my *xlt2 mur3-1* double mutant are XXFG subunits. In contrast, these subunits were not detected by MALDI-TOF MS. (However, because the number of XXFG subunits in the *xlt2 mur3-1* mutant appears to be affected by plant growth temperature (Kong *et al*, 2015), these data cannot be used to compare detection sensitivities with those of techniques used by other investigators.) Nevertheless, although PACE was able to distinguish isobaric structures such as XLXG and XXLG, MALDI-TOF MS was able to reveal the presence of xyloglucan subunits from *V. corymbosum* that were not identified by the former, most likely due to insufficient separation or coincidental co-migration. Particularly problematic co-migrations may include that of X☆XG and XXLG, and that of XXFG and X☆☆G. Hence, a combination of the two methods appears to be most effective. However, I found that the use of *exo*-glycosidases such as α-fucosidase and β-galactosidase substantially increased the power of both techniques, potentially permitting distinction between different monosaccharide substituents. The identification of new *exo*-glycosidases with better-characterised and more specific activities, such as the α-arabinofuranosidase *Bo*GH43A from *Bacteroides ovatus* (Larsbrink *et al*, 2014a), will likely be of benefit for these approaches.

Despite the limitations of these methods, I was able to characterise the activities of Cc07_g06550 and *Vm*GT47-A12[Δ167–193] with some confidence. Based on my data, Cc07_g06550 likely constitutes a β-xylosyltransferase to the second xylose in the XXXG repeat (when expressed in Arabidopsis), whereas *Vm*GT47-A12[Δ167–193] constitutes a weak α-arabinofuranosyltransferase to the third xylose in the XXXG repeat (**Figure 6.21**) (of course, the native acceptor substrate of these enzymes may not necessarily be based on an

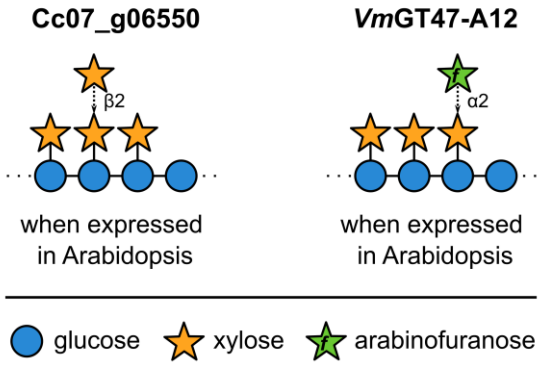


Figure 6.21 Most likely activities of Cc07_g06550 and *Vm*GT47-A12.

XXXG repeat). These conclusions are based on a number of lines of evidence; these are listed below along with their assumptions:

Firstly, PACE analysis clearly demonstrated that both enzymes were able to modify xyloglucan structure. However, the detection of these structures relies on the action of xyloglucanase: therefore, if the true substrate of either GT47-A enzyme is inaccessible to this xyloglucanase, the products will not be visible. Nevertheless, xyloglucanase has traditionally been assumed to recognise all xyloglucan in Arabidopsis cell walls.

Secondly, MALDI-TOF MS clearly demonstrated that the expression of these enzymes introduced at least one new subunit into the xyloglucan polysaccharide, appearing as an ion with a mass/charge ratio of 1217 m/z. However, although this ion almost certainly represents a sodiated H_4P_4 oligosaccharide, without further analysis by tandem MS (with collision-induced dissociation, CID), the linkages between these monosaccharides cannot be inferred from this information. Hence, I have assumed that this oligosaccharide, like the other major products, is built on the XXXG core, and that furthermore, these GT47-A enzymes transfer sugars to the C2 hydroxyls of the α-linked xylosyl residues (as other GT47-As do).

Thirdly, I used the α-arabinofuranosidase *Cj*Abf51 to reveal the presence of terminal α-Ara*f* residues on xyloglucan subunits introduced by *Sl*XST1 and *Vm*GT47-A12. This enzyme has previously been shown to cleave α1,3- and (to a lesser extent) α1,2-linked terminal Ara*f* residues from both xylan and arabinan (Beylot *et al*, 2001); hence, with such a broad specificity it should not be entirely surprising that it has some activity on arabinosylated xyloglucan. However, the activity of this enzyme on xyloglucan substrates was poor, and reactions did not proceed to completion. I was also unable to fully characterise potential side-activities/contaminating activities of this enzyme preparation, although it did not seem to exhibit activity on xyloglucan from *V. corymbosum*. Nonetheless, the enzyme was able to (partially) cleave Ara*f* residues introduced by *Sl*XST1, confirming that this enzyme preparation exhibits at least some α-arabinofuranosidase activity. Furthermore, in several replicates, and with two different enzyme preparations (data not shown), *Cj*Abf51 was unable to digest the xyloglucan subunit produced by Cc07_g06550. However, the possibility remains that this enzyme is unable to remove decorations from the second xylose in XXXG for some reason. Clearly, the use of a characterised xyloglucan α-arabinofuranosidase such as *Bo*GH43A (Larsbrink *et al*, 2014a) would have improved the power of these experiments; however, this enzyme was not available at the time of this work.

Fourthly, I used a GH3 β1,2-xylosidase from *C. globosum* to detect the presence of β-xylosylated xyloglucan subunits. This enzyme has only previously been shown to act on the terminal xylose in the $D^{2,3}$ structure found in grass xylan; hence, it was not expected to necessarily work on xyloglucan. Nevertheless, I was able to show that preparations of this enzyme were capable of removing pentose substituents from both *V. corymbosum* xyloglucan and the product of Cc07_g06550 activity, albeit at a slow rate. Curiously, a significant proportion of pentose substituents on the *V. corymbosum* xyloglucan were resistant to both this enzyme and *Cj*Abf51. As one possible explanation, I proposed that the *Cg*GH3 enzyme might only be able to remove β-Xyl residues from a particular position in the XXXG repeat (though the fact that it could remove β-Xyl from XUFG demonstrates that decorations at the second position are sensitive, at least). However, the main *Cg*GH3-resistant band from *V. corymbosum* xyloglucanase digests appeared to co-migrate with that of the *Cg*GH3-*sensitive* Cc07_g06550 product in PACE (**Figure 6.20**), suggesting that they are decorated at the same position. This implies that the difference in sensitivity is due to a difference in the identity of the pentose itself, rather than anything else. Nonetheless, due to its apparent β-galactosidase side activity/contamination and low activity on β-xylosylated xyloglucan, this enzyme does not represent an ideal tool for future structural analysis. The recently reported GH120 β1,2/4-xylosidase from *Herbinix* spp. strain LL1355 (Beri *et al*, 2020) could potentially provide an effective alternative; however, investigation of enzymes in *Bacteroides* spp. polysaccharide utilisation loci for xyloglucan digestion (Larsbrink *et al*, 2014a) might yield the most fruitful results.

Fifthly, I used an *E. coli* GH31 α-xylosidase and *An*GH3 β-glucosidase to deduce the positions of the decorations added by Cc07_g06550 and *Vm*GT47-A12. Initially, I showed that the α-xylosidase products GXLG and GLXG differ in their sensitivity to *An*GH3 (data not shown). The ability of this enzyme to cleave β-Glc from GXLG without cleaving that from GLXG is somewhat consistent with the previously reported activity of *Bo*GH3B β-glucosidase from *Bacteroides ovatus* (Larsbrink *et al*, 2014a). This enzyme appears to completely convert GXXG to XXG, whereas it has only partial success in converting GLLG to LLG. With this precedent, it is possible that *An*GH3 may also have some ability to digest GLXG over longer time periods or at higher enzyme concentrations. Nevertheless, I assumed that the difference in β-glucosidase sensitivity between GXLG and GLXG would be mirrored in the sensitivities of pentosylated oligosaccharides such as GXSG and GSXG. The observed sensitivity of GXSG was consistent with this prediction. Hence, it seems reasonable to conclude that the *An*GH3-

resistant α-xylosidase product from the Cc07_g06550 sample has the structure G☆XG (probably GUXG).

Ultimately, the products of these two enzymes will require characterisation with NMR. There was not enough time to achieve this during the course of the work, but in the future, purified, NMR-characterised oligosaccharides from plants over-expressing *Sl*XST1 or Cc07_g06550 could be used as standards for enzymatic assays.

The expression of either Cc07_g06550 or *Vm*GT47-A12 was able to complement the phenotypes of *mur3-3* and *xlt2 mur3-1*, albeit to different extents. The normal phenotypes of these mutants, especially that of *mur3-3*, are of particular interest because they are far more severe than the phenotype of plants lacking xyloglucan itself (Kong *et al*, 2015). The presence of endomembrane aggregates in *mur3-3* mutants, as well as the fact that the *mur3-3* phenotype can be rescued by high temperature, has led to the hypothesis that the phenotype is caused by aggregation of xyloglucan in the Golgi due to its low level of galactosylation (Tamura *et al*, 2005; Kong *et al*, 2015). The fact that the single mutant *mur3-3* has a more severe phenotype than the double mutant *xlt2 mur3-1* (Tamura *et al*, 2005; Jensen *et al*, 2012; Kong *et al*, 2015) could indicate that the few fucosylated sidechains of the latter are better able to prevent such aggregation than the second-position galactosyl sidechains of the former (as fucosylation can only occur on Gal residues added by MUR3). However, the *fut1* mutant does not exhibit a phenotype by itself, at least.

The functional equivalency between galactosyl and arabinosyl sidechains has already been demonstrated to some extent by the fact that expression of *Sl*XST1/2 or *Pp*XDT complements the phenotype of the *xlt2 mur3-1* mutant (Schultink *et al*, 2013; Zhu *et al*, 2018). However, *xlt2 mur3-1*, unlike *mur3-3*, still contains a small but significant amount of fucosylation. Therefore, combined with the previous finding that the *mur3-3* phenotype can be rescued by overexpression of XLT2 (Kong *et al*, 2015), my finding that the expression of pentosyltransferases Cc07_g06550 and *Vm*GT47-A12 is also able to rescue *mur3-3* demonstrates clearly for the first time that neither galactosylation nor fucosylation is required at the third position in the XXXG repeat for normal growth in Arabidopsis. Even so, taken as a whole, these results hint at a fine threshold in the level of xyloglucan decoration under which deleterious effects occur, and against which fucosylgalactosyl-xylose trisaccharide sidechains are a better defence than disaccharide sidechains. Consistent with this hypothesis is the fact that, in my experiments, the level of complementation did not appear to be proportional to the

level of activity from the GT47-A enzymes: for instance, expression of *Vm*GT47-A12[Δ167–193] under the XXT2 promoter resulted in a consistent partial, yet substantial complementation effect in *mur3-3*, despite a virtually undetectable level of xyloglucan modification. Interestingly, though, the position of the decorations could still have a role in ameliorating these phenotypes, as plants over-expressing *Sl*XST1 appeared to grow very slightly taller than those over-expressing Cc07_g06550. This could be due to the fact that *Sl*XST1 was able to introduce a higher level of doubly pentosylated subunits into the xyloglucan structure (*i.e.* XSSG); however, it could also possibly due to unintentional differences in abiotic factors between the two trays of plants.

The reason for which *Vm*GT47-A12 displayed such a low level of activity was not entirely clear, but may have been caused by the perhaps unwise decision to remove amino acids 167–193 from the stem domain. Alternatively, this enzyme may simply be degenerate—the recent duplications of XST homologues in *Vaccinium* spp. and *Rhododendron* spp. may indeed have created some redundant copies that have already lost some activity. The anomalous residues in the Nβ5–Nα5 loop in these enzymes may therefore reflect a loss of activity, rather than the gain of a new one. This could also be true for Bv7_174350_exki and Cc07_g06570, although it is also possible that they were not expressed to sufficient levels, or that Arabidopsis does not provide the correct environment to see their activity. Nonetheless, the potential existence of such degenerate enzymes will pose an obstacle to the prediction of new GT47-A activities based on sequence. Furthermore, if *Vm*GT47-A12 is a duplicate arabinofuranosyltransferase, the question is raised as to which other enzyme is responsible for creating the 'U' sidechains seen in *V. macrocarpon*. A larger *mur3-3*-based screening approach may be necessary to answer this question; however, it may be easier to look instead at *Argania spinosa* GT47-A enzymes, as they are fewer in number.

*Vm*GT47-A12 does not appear to constitute the β-xylosyltransferase it was speculated to be. However, in spite of this, I discovered that the XST2 homologue Cc07_g06550 appears to exhibit this very activity. This result is surprising because β-xylosyl-xylose disaccharide sidechains have not previously been reported outside the Ericales order. However, innovations such as these are likely to have occurred multiple times in the evolution of plants, as they likely provide (at least temporary) protection from the glycosidases secreted by invasive micro-organisms. Unfortunately, due to the scarcity of xyloglucan in *C. arabica* leaf tissue, I have not yet been able to confirm the presence of this decoration in native cell walls. Nonetheless, the fact that this enzyme was identified because it exhibited unusual residues in its Nβ5–Nα5 loop

provides the proof of concept that I set out to obtain. Whether this particular example is coincidental is hard to gauge; nevertheless, my bioinformatics results regarding the whole GT47-A clade suggest that there is an effect at play across the whole family. Glycosyltransferase nucleotide sugar specificities have been consistently difficult to understand from a rational perspective—nonetheless, my results suggest that the GT47-A clade could constitute a valuable model system in this pursuit. It seems likely that the advent of machine-learning techniques will help to make sense of the subtle changes that create new activities in these fascinating enzymes.

Chapter 7 : Conclusions and future work

In 1958, Francis Crick published his ‘sequence hypothesis’: the revolutionary idea that information is transferred from nucleic acid sequence to protein sequence (Crick, 1958, 1970). By the end of that same year, the first protein crystal structure had been published (Kendrew *et al*, 1958), and just three years later, pioneering experiments on ribonuclease A folding had been conducted (White, 1961; Anfinsen *et al*, 1961). The latter would soon lead to the development of ‘Anfinsen’s dogma’: simply put, that the native conformation of small, globular proteins is determined by their amino acid sequence (Anfinsen, 1973). In the sixty years since, the field of molecular biology has, of course, become more complex; nevertheless, the idea that nucleic acid sequence determines protein sequence, and therefore largely protein structure, remains a central aspect in biology—an idea embodied in the recent breakthrough in computational structure prediction by AlphaFold (Callaway, 2020). Hence, the prediction of protein structure and function from nucleic acid or protein sequence has a strong basis in established theory.

Accordingly, the work presented in this thesis constitutes an attempt to rationalise and predict the structures and functions of glycosyltransferases from families GT43, GT47, and GT64 using their (nucleic acid-derived) protein sequences. The outcomes of this work therefore not only shed insight on the biology of these GTs, but also lay down some groundwork in the pursuit of structures and functions for other, uncharacterised GTs. Broadly speaking, my results concern two major aspects of GT structure-function relationships: Golgi GT homodimerisation, and the structural basis for nucleotide sugar recognition.

7.1 Golgi glycosyltransferase homodimerisation

In this work, I have provided evidence to strengthen the idea that many Golgi glycosyltransferases form symmetric homodimers. For example, my results suggest that the TMHs of plant GT43 members IRX9 and IRX14 undergo GAS_{right} motif-mediated homodimerisation (*Sections 3.2.2–3*). I showed that certain residues in the TMHs of these proteins are highly conserved, and that when expressed in isolation in *E. coli*, these TMH peptides appear to self-associate. Furthermore, I showed that Gly28 is necessary for homo-oligomerisation (presumably homodimerisation) of the $IRX9_{21–34}$ TMH fragment in the *E. coli* inner membrane. In addition, I showed that Gly28, but not Cys24, appears to be necessary for the function of full-length IRX9 in Arabidopsis, and that mutation of this residue may result in

altered expression or localisation (*Sections 3.2.4–5*). Precedent for such homodimerisation (or homo-oligomerisation) exists in previous reports of Golgi GT TMHs linked by disulphide bridges (Tu & Banfield, 2010). However, the homodimerisation of the IRX9 TMH appears to be the first amongst Golgi GTs to be shown to be mediated by a specific amino acid motif. The fact that small and/or polar residues are relatively common in the TMHs of other Golgi glycosyltransferases (Sharpe *et al*, 2010; Tu & Banfield, 2010) suggests that interactions of this type could be widespread in the Golgi proteome. Indeed, as mentioned in **Section 3.3**, the lipophobic effect may push TMHs together in the membrane; hence, it is easy to see how stabilising features or interaction surfaces might have evolved over time (though equally this effect may render residue-mediated homodimerisation redundant in some cases). Therefore, TMHs from other Golgi GTs should be tested for oligomerisation using TOXGREEN or similar techniques. Computational techniques for predicting TMH dimerisation such as CATM (Mueller *et al*, 2014) and THOIPA (Xiao *et al*, 2020) may also be useful for analysing large sequence datasets. Furthermore, around twenty TMH dimer structures have been successfully solved by NMR or X-ray crystallography (Bugge *et al*, 2016; Bocharov *et al*, 2017; Valley *et al*, 2017); hence, it might also be possible to solve the structure of the IRX9 and IRX14 dimers (this would likely involve expression in *E. coli*). Such a structure would constitute the first of any canonical glycosyltransferase CTS domain.

By identifying key interface residues in human GT43 crystal structures and comparing them to aligned residues in plant GT43s, I argued that the globular domains of plant GT43s probably also form symmetrical homodimers (*Section 3.2.6*). The fact that GT43 homodimerisation is shared across the two kingdoms suggests that it is an ancient characteristic, predating xylan synthesis. Hence, it seems unlikely that IRX9 homodimerisation is a specific mechanism for 'xylan synthase complex' assembly (though it may still be necessary for this process). Conversely, although homodimerisation of the human GT43s introduces new sidechains into the proximity of the active site, IRX9 is not thought to be active—hence, GT43 homodimerisation does not seem as though it could have a universally important role in reshaping the active site. More likely, therefore, is the notion that homodimerisation has a more fundamental role in Golgi GT biology.

The most closely related family to GT43 is currently thought to be the GT64 family (Taujale *et al*, 2020). In this work, I used cryo-EM to confirm that the GT64 domains in EXTL3 and EXTL2 form *bona fide* symmetrical homodimers (previously, homodimeric GT conformations could only be confirmed by virtue of the presence of intermolecular disulphide bonds, which

are not present in *Mm*EXTL2; *Sections 4.2.4–5*). The EXTL3 homodimer in particular appears to be stabilised by a pair of disulphide bridges in its GT64 domains as well as a predicted coiled-coil domain in its stem region. These stabilisation measures suggest that homodimerisation has an important role in EXTL3 function. Indeed, I proposed that the C-terminus of each GT64 domain might aid acceptor binding in the active site of the opposing monomer—which could explain the importance of homodimerisation in this enzyme. However, at current, this hypothesis is speculative. Furthermore, comparing catalytic domain homodimerisation between GT64 and GT43 enzymes may, in fact, shed more insight: the involvement of both hypervariable region 2 and the post-β7' loop in the dimerisation of GTs from both families suggests that homodimerisation might have been inherited from their common ancestor. Indeed, these structural elements have also been found to participate in homodimeric interfaces in enzymes from other GT-A families (Harrus *et al*, 2018). However, the actual orientation of the two GT-A folds with respect to each other is very different between GT43 and GT64—with α1 and the active sites at, or close to, the interface in GT43s, but separated by a considerable distance in GT64s. Hence, it seems that the orientation of the two active sites with respect to one another is unlikely to be a well conserved aspect in GT homodimerisation; therefore, the fundamental role of homodimerisation seems unlikely to lie in catalysis. Further investigation into the role of homodimerisation in localisation, as discussed in **Chapter 3**, will likely shed light on this matter.

7.2 Nucleotide sugar binding

In this work, I also reported the structure of the EXTL3 GT47 domain, confirming that enzymes in the GT47 family adopt a GT-B fold (*Section 4.2.6*). Furthermore, I reported the presence of the 'glycogen phosphorylase-like' motif (Wrabl & Grishin, 2001) within this family—including, in particular, a conserved Asp/Glu residue in the Cα4 helix that likely binds the nucleotide sugar ribose moiety in these enzymes. Although I found that several previously characterised inactivating mutations can be mapped to the vicinity of this helix in human and Chinese hamster EXT1, I reasoned that this helix does not have a role in the activity of EXTL3. This was because the potential nucleotide binding pocket appears to have been obstructed by two recent modifications: the extension of Cα4 and the formation of a salt bridge between Arg421 and Glu453. Furthermore, whereas the EXTL3 GT64 active site could clearly bind UDP, I could not obtain a structure with UDP bound in the GT47 domain active site (*Section 4.2.7*). However, these observations were at odds with the results of my activity assays, which indicated the unambiguous presence of a GlcAT-II activity in EXTL3 preparations that could

not be fully inhibited by EDTA (*Sections 4.2.2–3*)—creating somewhat of a paradox which requires urgent resolution. The true role of EXTL3ΔN in the observed GlcAT-II and GlcNAcT-II activities could be tested in the future by expressing and purifying EXTL3ΔN with relevant inactivating point mutations.

Nevertheless, I was able to apply the knowledge gained from the EXTL3 GT47 structure to several questions concerning related enzymes in plants. This enabled several new experiments. For example, I tried to rationalise the difference in nucleotide sugar specificity between related XAPT and XLPT enzymes, which belong to GT47-A. However, although a single (Ala→Gly) residue change located in the predicted Nβ5–Nα5 loop initially appeared to represent a promising explanation for the development of galactosyltransferase activity (*Section 5.2.3*), I also showed that this amino acid change is neither sufficient nor necessary for the change in activity (*Sections 5.2.4–6*). In the absence of any other amino acid changes that could convincingly explain this activity change, I concluded that the nucleotide sugar specificity of these enzymes may not be determined simply by the presence or absence of particular sidechains surrounding the binding pocket. This could be tested in the future by assaying the activities of further point mutants and chimeras.

In spite of this hypothesis, I proposed that amino acid identities within the Nβ5–Nα5 loop are nonetheless linked to (and perhaps somewhat indicative of) the nucleotide sugar specificities of GT47-A enzymes. In this work, I provided evidence for this idea by demonstrating a correlation between Nβ5–Nα5 loop residues and donor specificity amongst different GT47-A clusters (*Section 6.2.1*). Furthermore, I demonstrated that several previously uncharacterised GT47-A enzymes with anomalous residues in this loop exhibit nullified, reduced, or novel activities. In particular, I found that a *C. canephora* XST homologue Cc07_g06550, which possesses unusual Nβ5–Nα5 loop residues, exhibits a novel activity on xyloglucan, and perhaps constitutes a β-xylosyltransferase (*Sections 6.2.3–9*). The activity of this enzyme might be confirmed in the future by better analysing its product—perhaps by NMR, MS/MS with CID, and/or digestion with better-characterised glycosyl hydrolases. Future work regarding nucleotide sugar specificity in GT47-A might involve investigating the activities of other enzymes with anomalous Nβ5–Nα5 loop residues or with further point mutations in this loop. It will also be interesting to see whether these ideas can be extrapolated to other GT47 clades.

All the same, my results are consistent with the idea that the specification of donor preference within GT47-A enzymes is subtle and not easily predicted by rational approaches. Of course,

future structures of these enzymes will help confirm or rebut this hypothesis; nevertheless, the sheer diversity of donor specificities within this clade suggests that such specificity is easily perturbed—perhaps by changes distal to the active site. Hence, hopefully with a more comprehensive knowledge of the donor specificities of GT47-A members, machine learning techniques may be required to detect the true principles and patterns that relate sequence, structure, and activity in this family.

7.3 Golgi glycosyltransferase fidelity and hetero-complex formation

It is becoming increasingly apparent that many Golgi GTs exhibit some level of promiscuity in their activity (Biswas & Thattai, 2020). Although this can often simply constitute an indifference to minor variations in acceptor structure, several Golgi GTs are in fact known to exhibit promiscuity in their nucleotide sugar specificity. The human enzyme β4GalT1 provides a particularly pertinent example—it is thought that this enzyme has, in addition to its preferred galactosyltransferase activity, low levels of glucosyltransferase and *N*-acetylgalactosaminyltransferase activity; furthermore, these activities can be stimulated by the binding of lactalbumin (Biswas & Thattai, 2020). Similarly, the Arabidopsis β-galactosyltransferase GALS1 has been reported to possess a secondary α-arabinopyranosyltransferase activity that can be stimulated by increased UDP-Ara*p* concentration (Laursen *et al*, 2018). Furthermore, in the Arabidopsis *mur1* mutant, which cannot import GDP-fucose into the Golgi, the fucosyltransferase FUT1 is able to utilise GDP-L-galactose in place of GDP-fucose, creating L-galactosylated xyloglucan (Zablackis *et al*, 1996; Ohashi *et al*, 2018). While in many cases, such multifunctionality could perform a useful role (by performing two reactions with one enzyme), it is easy to imagine that unbridled promiscuity could create serious problems in other glycosylation pathways—for instance when the precise structure of the synthesised glycan has an 'informational' role (*e.g.* in cell–cell communication). Hence, it seems plausible that some GTs normally benefit from additional measures to ensure their proper fidelity. Indeed, the fact that different cell types and species are able to generate distinct and well defined glycan profiles despite possessing the same complement of promiscuous Golgi GTs has led to the hypothesis that the activities of these enzymes are controlled by strict compartmentalisation (Jaiman & Thattai, 2020).

In reality, however, individual GTs are not thought to be strictly confined to particular cisternæ, but rather are thought to be distributed in gradients across the Golgi stacks (Young, 2004). Nevertheless, it is still possible that GTs could be compartmentalised in a sub-cisternal manner. Indeed, the existence of several bi-domain glycosyltransferases in eukaryote proteomes (such

as exostosins, chondroitin synthases, and LARGE1/2 in humans) suggests that the physical association of related GT activities is of benefit to the cell. In this work, I reported the domain architecture of bi-domain exostosin EXTL3, from which the likely domain architecture of the other exostosins can be inferred. The results did not suggest that the two alternating activities of these enzymes are mechanistically linked, but rather that the formation of bi-domain GTs likely constitutes a simple means to increase local acceptor substrate concentrations (*Section 4.2.7*). Furthermore, I also reported the existence of a predicted tri-domain GT present in lower plants, with a GT64-GT8R-GT8R domain structure (*Section 4.2.10*). Hence, multi-domain GTs appear to have emerged multiple times in the eukaryotes as a result of convergent evolution—supporting the idea that glycosyltransferase activities can universally benefit from the formation of tailored 'nano-environments'.

A similar effect is thought to be achieved through the non-covalent formation of hetero-complexes between GTs in the same pathway (Young, 2004; de Graffenried & Bertozzi, 2004; Oikawa *et al*, 2013). This phenomenon is well documented and has been frequently supported by co-immunoprecipitation and FRET experiments (Young, 2004; de Graffenried & Bertozzi, 2004; Kellokumpu *et al*, 2016); however, no structure of such a complex has yet been published (Khoder-Agha *et al*, 2019a). In addition, there has not always been a clear distinction in the literature between the formation of strictly heterodimeric complexes (which, when occurring between two highly related GTs such as EXT1 and EXT2, could potentially arise from pseudosymmetric interactions) and the interaction of existing GT homodimers to form higher-order oligomers. It has been proposed that homodimers dissociate as they progress through the Golgi in order to form hetero-complexes (Kellokumpu *et al*, 2016). However, how the intermolecular disulphide bonds that stabilise many of these homodimers could be broken to achieve this has not been discussed. Furthermore, it has been proposed that the xylan synthase complex in asparagus contains homodimers of IRX9, IRX14, and IRX10 orthologues (Zeng *et al*, 2016). Indeed, the fact that I found conserved cysteine residues in the TMHs of these enzymes suggests that these dimers could be covalently linked (*Section 3.2.1*). It would therefore be of interest to analyse the plant GT43s using SDS-PAGE under reducing and non-reducing conditions to confirm whether these homodimers are indeed disulphide-bonded.

In any case, it is unclear exactly how homodimeric GTs might form higher-order oligomers in the narrow Golgi lumen. If conventional protein interfaces are involved, it is possible that the symmetry of each homodimer would drive the formation of large aggregates or arrays of GTs within the membrane—somewhat similarly to that envisaged in the early kin recognition model

(Nilsson *et al*, 1993). Indeed, repeating array structures have been observed at the centre of cisternæ in cryo-EM tomograms of the *C. reinhardtii* Golgi (Engel *et al*, 2015). However, given the role of the CTS domain in overall Golgi localisation, it is possible that transmembrane domains could have some responsibility for co-localising GTs—be it through precise targeting to specific lipid domains within each cisterna, or by the direct association of TMHs within the membrane. It is also yet to be ruled out that cytoplasmic factors could somehow organise the sub-cisternal localisation of GTs by acting on their cytoplasmic tails. In addition to such organisation within the plane of the membrane, there is reason to believe that multiple nano-environments could also exist in the membrane-lumen axis: the coiled coil and noncatalytic GT47 domain of EXTL3 suggest that the EXTL3 GT64 domain benefits from being located distally from the membrane, for example. Therefore, there exists an interesting potential for three-dimensional, nano-scale organisation within each cisterna. The existence of such a phenomenon would help to explain how the fidelity of glycosyltransferase reactions is maintained and regulated. These ideas may be tested in the future by higher resolution tomograms of the Golgi apparatus.

7.4 Concluding remarks

Ultimately, when combined, Golgi GT activity, localisation, and oligomerisation could well be determined almost entirely by protein sequence—permitting control of the cell surface glycome at a genomic level. In the work presented above, this principle has been reflected in three broad concepts: the role of transmembrane and catalytic domain residues in the homodimerisation and localisation of GT43 members, the role of Nβ5–Nα5 and Cα4 residues in nucleotide sugar binding by GT47 members, and the role of various catalytic domain residues in the activity and homodimerisation of GT64 members. Because many Golgi GTs share similarities in their architecture and catalytic-domain folds, the sequence-structure-function relationships exhibited by these GT43-, GT47-, and GT64-family enzymes can potentially be extrapolated to Golgi GTs from other families, facilitating the prediction of protein behaviour from sequence and the design of new experiments to investigate Golgi glycosylation machinery. Combined with ever-improving computational techniques to predict protein structure and function (Yang *et al*, 2018; Taujale *et al*, 2020; Callaway, 2020), these concepts will hopefully lay down a path towards a more comprehensive understanding of Golgi glycosyltransferase behaviour, as well as an improved understanding of the functioning of the Golgi apparatus itself.

References

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, Cárdenas P, Čepička I, Chistyakova L, del Campo J, Dunthorn M, Edvardsen B, Eglit Y, Guillou L, Hampl V, Heiss AA, et al (2019) Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J. Eukaryot. Microbiol.* **66:** 4–119

Aebi M (2013) N-linked protein glycosylation in the ER. *Biochim. Biophys. Acta - Mol. Cell Res.* **1833:** 2430–2437

Afonine P V., Headd JJ, Terwilliger TC & Adams PD (2013) New tool: phenix.real_space_refine. *Comput. Crystallogr. Newsl.* **4:** 43–44

Ahnert SE, Marsh JA, Hernandez H, Robinson C V. & Teichmann SA (2015) Principles of assembly reveal a periodic table of protein complexes. *Science* **350:** aaa2245

Albesa-Jové D, Cifuente JO, Trastoy B & Guerin ME (2019) Quick-soaking of crystals reveals unprecedented insights into the catalytic mechanism of glycosyltransferases. *Methods Enzymol.* **621:** 261–279

Albesa-Jové D, Giganti D, Jackson M, Alzari PM & Guerin ME (2014) Structure-function relationships of membrane-associated GT-B glycosyltransferases. *Glycobiology* **24:** 108–124

Allen HJ & Kisailus EC (1992) Glycoconjugates: composition, structure, and function. New York: Dekker

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402

Amos RA & Mohnen D (2019) Critical review of plant cell wall matrix polysaccharide glycosyltransferase activities verified by heterologous protein expression. *Front. Plant Sci.* **10:** 915

Anders N & Dupree P (2011) Glycosyltransferases of the GT43 Family. In *Annual Plant Reviews, Volume 41: Plant Polysaccharides* pp 251–263. Chichester: Wiley-Blackwell

Anderson SM, Mueller BK, Lange EJ & Senes A (2017) Combination of Cα-H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes. *J. Am. Chem. Soc.* **139:** 15774–15783

Anderson SM & Senes A (2018) Email to Louis Wilson.

References

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* **181:** 223–230

Anfinsen CB, Haber E, Sela M & White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **47:** 1309–1314

Aoki D, Lee N, Yamaguchi N, Dubois C & Fukuda MN (1992) Golgi retention of a trans-Golgi membrane protein, galactosyltransferase, requires cysteine and histidine residues within the membrane-anchoring domain. *Proc. Natl. Acad. Sci. U. S. A.* **89:** 4319–4323

Aquino RS, Grativol C & Mourão PAS (2011) Rising from the Sea: Correlations between Sulfated Polysaccharides and Salinity in Plants. *PLoS One* **6:** e18862

Ardèvol A, Iglesias-Fernández J, Rojas-Cervellera V & Rovira C (2016) The reaction mechanism of retaining glycosyltransferases. *Biochem. Soc. Trans.* **44:** 51–60

Armstrong CR & Senes A (2016) Screening for transmembrane association in divisome proteins using TOXGREEN, a high-throughput variant of the TOXCAT assay. *Biochim. Biophys. Acta - Biomembr.* **1858:** 2573–2583

Atmodjo MA, Hao Z & Mohnen D (2013) Evolving Views of Pectin Biosynthesis. *Annu. Rev. Plant Biol.* **64:** 747–779

Atmodjo MA, Sakuragi Y, Zhu X, Burrell AJ, Mohanty SS, Atwood JA, Orlando R, Scheller H V, Mohnen D & Mohnen D (2011) Galacturonosyltransferase (GAUT)1 and GAUT7 are the core of a plant cell wall pectin biosynthetic homogalacturonan:galacturonosyltransferase complex. *Proc. Natl. Acad. Sci. U. S. A.* **108:** 20225–20230

Awad W, Kjellström S, Svensson Birkedal G, Mani K & Logan DT (2018) Structural and Biophysical Characterization of Human EXTL3: Domain Organization, Glycosylation, and Solution Structure. *Biochemistry* **57:** 1166–1177

Bahaji A, Li J, Sánchez-López ÁM, Baroja-Fernández E, Muñoz FJ, Ovecka M, Almagro G, Montero M, Ezquer I, Etxeberria E & Pozueta-Romero J (2014) Starch biosynthesis, its regulation and biotechnological approaches to improve crop yields. *Biotechnol. Adv.* **32:** 87–106

Bard J (2017) Principles of evolution: systems, species, and the history of life. London: Garland Science

Beck R, Ravet M, Wieland FT & Cassel D (2009) The COPI system: Molecular mechanisms and function. *FEBS Lett.* **583:** 2701–2709

Becker JL, Tran DT & Tabak LA (2018) Members of the GalNAc-T family of enzymes utilize

distinct Golgi localization mechanisms. *Glycobiology* **28:** 841–848

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van De Peer Y, Coppens F & Vandepoele K (2018) PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* **46:** D1190–D1196

Berg JM, Tymoczko JL, Stryer L & Stryer L (2002) Biochemistry 5th ed. New York: W.H. Freeman

Beri D, York WS, Lynd LR, Peña MJ & Herring CD (2020) Development of a thermophilic coculture for corn fiber conversion to ethanol. *Nat. Commun.* **11:** 1937

Bernfield M, Götte M, Park PW, Reizes O, Fitzgerald ML, Lincecum J & Zako M (1999) Functions of Cell Surface Heparan Sulfate Proteoglycans. *Annu. Rev. Biochem.* **68:** 729–777

Beylot M-H, McKie VA, Voragen AGJ, Doeswijk-Voragen CHL & Gilbert HJ (2001) The Pseudomonas cellulosa glycoside hydrolase family 51 arabinofuranosidase exhibits wide substrate specificity. *Biochem. J.* **358:** 607–614

Bian Y, Ballington J, Raja A, Brouwer C, Reid R, Burke M, Wang X, Rowland LJ, Bassil N & Brown A (2014) Patterns of simple sequence repeats in cultivated blueberries (Vaccinium section Cyanococcus spp.) and their use in revealing genetic diversity and population structure. *Mol. Breed.* **34:** 675–689

Biswas A & Thattai M (2020) Promiscuity and specificity of eukaryotic glycosyltransferases. *Biochem. Soc. Trans.* **48:** 891–900

Bocharov E V., Mineev KS, Pavlov K V., Akimov SA, Kuznetsov AS, Efremov RG & Arseniev AS (2017) Helix-helix interactions in membrane domains of bitopic proteins: Specificity and role of lipid environment. *Biochim. Biophys. Acta - Biomembr.* **1859:** 561–576

Bock KW (2016) The UDP-glycosyltransferase (UGT) superfamily expressed in humans, insects and plants: Animal-plant arms-race and co-evolution. *Biochem. Pharmacol.* **99:** 11–17

Borner GHH, Sherrier DJ, Weimar T, Michaelson L V., Hawkins ND, MacAskill A, Napier JA, Beale MH, Lilley KS & Dupree P (2005) Analysis of detergent-resistant membranes in arabidopsis. Evidence for plasma membrane lipid rafts. *Plant Physiol.* **137:** 104–116

Brandon AG, Birdseye DS & Scheller H V. (2020) A dominant negative approach to reduce xylan in plants. *Plant Biotechnol. J.* **18:** 5–7

Breton C, Fournel-Gigleux S & Palcic MM (2012) Recent structures, evolution and mechanisms of glycosyltransferases. *Curr. Opin. Struct. Biol.* **22:** 540–549

References

Breton C, Mucha J & Jeanneau C (2001) Structural and functional features of glycosyltransferases. *Biochimie* **83:** 713–718

Breton C, Najdrová LŠ, Jeanneau C, Koca J & Imberty A (2006) Structures and mechanisms of glycosyltransferases. *Glycobiology* **16:** 29–37

Bretscher MS & Munro S (1993) Cholesterol and the Golgi Apparatus. *Science* **261:** 1280–1281

Briggs DC & Hohenester E (2018) Structural Basis for the Initiation of Glycosaminoglycan Biosynthesis by Human Xylosyltransferase 1. *Structure* **26:** 801-809.e3

Brodribb TJ, Carriquí M, Delzon S, McAdam SAM & Holbrook NM (2020) Advanced vascular function discovered in a widespread moss. *Nat. Plants* **6:** 273–279

Bromley JR, Busse-Wicher M, Tryfona T, Mortimer JC, Zhang Z, Brown DM & Dupree P (2013) GUX1 and GUX2 glucuronyltransferases decorate distinct domains of glucuronoxylan with different substitution patterns. *Plant J.* **74:** 423–434

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P & Turner SR (2007) Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. *Plant J.* **52:** 1154–1168

Brown DM, Zeef LAH, Ellis J, Goodacre R & Turner SR (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17:** 2281–2295

Brown DM, Zhang Z, Stephens E, Dupree P & Turner SR (2009) Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in Arabidopsis. *Plant J.* **57:** 732–746

Brunet T & King N (2017) The Origin of Animal Multicellularity and Cell Differentiation. *Dev. Cell* **43:** 124–140

Bugge K, Lindorff-Larsen K & Kragelund BB (2016) Understanding single-pass transmembrane receptor signaling from a structural viewpoint—what are we missing? *FEBS J.* **283:** 4424–4451

Bülter T & Elling L (1999) Enzymatic synthesis of nucleotide sugars. *Glycoconj. J.* **16:** 147–159

Burki F, Roger AJ, Brown MW & Simpson AGB (2020) The New Tree of Eukaryotes. *Trends Ecol. Evol.* **35:** 43–55

Burnley T, Palmer CM & Winn M (2017) Recent developments in the CCP-EM software suite. *Acta Crystallogr. Sect. D Struct. Biol.* **73:** 469–477

Burton RA & Fincher GB (2012) Current challenges in cell wall biology in the cereals and

grasses. *Front. Plant Sci.* **3:** 130

Busse-Wicher M, Li A, Silveira RL, Pereira CS, Tryfona T, Gomes TCF, Skaf MS & Dupree P (2016) Evolution of xylan substitution patterns in gymnosperms and angiosperms: Implications for xylan interaction with cellulose. *Plant Physiol.* **171:** 2418–2431

Busse-Wicher M, Wicher KB & Kusche-Gullberg M (2014) The exostosin family: Proteins with many functions. *Matrix Biol.* **35:** 25–33

Busse M, Feta A, Presto J, Wilén M, Grønning M, Kjellén L & Kusche-Gullberg M (2007) Contribution of EXT1, EXT2, and EXTL3 to heparan sulfate chain elongation. *J. Biol. Chem.* **282:** 32802–10

Busse M & Kusche-Gullberg M (2003) In vitro polymerization of heparan sulfate backbone by the EXT proteins. *J. Biol. Chem.* **278:** 41333–41337

Cacas JL, Buré C, Grosjean K, Gerbeau-Pissot P, Lherminier J, Rombouts Y, Maes E, Bossard C, Gronnier J, Furt F, Fouillen L, Germain V, Bayer E, Cluzet S, Robert F, Schmitter JM, Deleu M, Lins L, Simon-Plas F & Mongrand S (2016) Revisiting plant plasma membrane lipids in tobacco: A focus on sphingolipids. *Plant Physiol.* **170:** 367–384

Calder PC (1991) Glycogen structure and biogenesis. *Int. J. Biochem.* **23:** 1335–1352

Callaway E (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588:** 203–204

Campbell JA, Davies GJ, Bulone V & Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem. J.* **326:** 929–939

Carpita NC & Gibeaut DM (1993) Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J.* **3:** 1–30

Cartmell A, Lowe EC, Baslé A, Firbank SJ, Ndeh DA, Murray H, Terrapon N, Lombard V, Henrissat B, Turnbull JE, Czjzek M, Gilbert HJ & Bolam DN (2017) How members of the human gut microbiota overcome the sulfation problem posed by glycosaminoglycans. *Proc. Natl. Acad. Sci. U. S. A.* **114:** 7037–7042

Cavalier-Smith T (2017) Origin of animal multicellularity: Precursors, causes, consequences—the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. *Philos. Trans. R. Soc. B Biol. Sci.* **372:** 20150476

Chang A, Singh S, Phillips GN & Thorson JS (2011) Glycosyltransferase structural biology and its role in the design of catalysts for glycosylation. *Curr. Opin. Biotechnol.* **22:** 800–808

References

Charollais J & Van Der Goot FG (2009) Palmitoylation of membrane proteins (Review). *Mol. Membr. Biol.* **26:** 55–66

Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Q, Yang L, Sheng Y, Zhou Y, Xu H, Xie H, Long X, Zhang J, Wang Z, Shi M, Lu Y, Liu S, Guan L, Zhu Q, et al (2019) Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nat. Plants* **5:** 18–25

Chen X, Vega-Sánchez ME, Verhertbruggen Y, Chiniquy D, Canlas PE, Fagerström A, Prak L, Christensen U, Oikawa A, Chern M, Zuo S, Lin F, Auer M, Willats WGT, Bartley L, Harholt J, Scheller H V. & Ronald PC (2013) Inactivation of OsIRX10 leads to decreased xylan content in rice culm cell walls and improved biomass saccharification. *Mol. Plant* **6:** 570–573

Chen YH, Narimatsu Y, Clausen TM, Gomes C, Karlsson R, Steentoft C, Spliid CB, Gustavsson T, Salanti A, Persson A, Malmström A, Willén D, Ellervik U, Bennett EP, Mao Y, Clausen H & Yang Z (2018) The GAGOme: a cell-based library of displayed glycosaminoglycans. *Nat. Methods* **15:** 881–888

Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, Sun W, Li X, Xu Y, Zhang Y, Wittek S, Reder T, Günther G, Gontcharov A, Wang S, Li L, Liu X, Wang J, Yang H, Xu X, et al (2019) Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179:** 1057–1067

Chou Y-H, Pogorelko G & Zabotina OA (2012) Xyloglucan xylosyltransferases XXT1, XXT2, and XXT5 and the glucan synthase CSLC4 form Golgi-localized multiprotein complexes. *Plant Physiol.* **159:** 1355–1366

Chumpen Ramirez S, Ruggiero FM, Daniotti JL & Valdez Taubas J (2017) Ganglioside glycosyltransferases are S-acylated at conserved cysteine residues involved in homodimerisation. *Biochem. J.* **474:** 2803–2816

Clough SJ & Bent AF (1998) Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J.* **16:** 735–743

Colley KJ (1997) Golgi localization of glycosyltransferases: more questions than answers

Corfield AP (2015) Mucins: A biologically relevant glycan barrier in mucosal protection. *Biochim. Biophys. Acta - Gen. Subj.* **1850:** 236–252

Cosgrove DJ (2014) Re-constructing our models of cellulose and primary cell wall assembly. *Curr. Opin. Plant Biol.* **22:** 122–131

Cosgrove DJ & Jarvis MC (2012) Comparative structure and biomechanics of plant primary and secondary cell walls. *Front. Plant Sci.* **3:** 204

Couchman JR & Pataki CA (2012) An Introduction to Proteoglycans and Their Localization. *J. Histochem. Cytochem.* **60:** 885–897

Coutinho PM, Deleury E, Davies GJ & Henrissat B (2003) An evolving hierarchical family classification for glycosyltransferases. *J. Mol. Biol.* **328:** 307–317

Crick FH (1958) On protein synthesis. *Symp. Soc. Exp. Biol.* **12:** 138–163

Crick FH (1970) Central dogma of molecular biology. *Nature* **227:** 561–563

Croll TI (2018) ISOLDE: A physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. Sect. D Struct. Biol.* **74:** 519–530

Crooks GE, Hon G, Chandonia J-M & Brenner SE (2004) WebLogo: A Sequence Logo Generator. *Genome Res.* **14:** 1188–1190

Culbertson AT, Chou Y-H, Smith AL, Young ZT, Tietze AA, Cottaz S, Fauré R & Zabotina OA (2016) Enzymatic Activity of Xyloglucan Xylosyltransferase 5. *Plant Physiol.* **171:** 1893–1904

D'Arienzo A, Andreani L, Sacchetti F, Colangeli S & Capanna R (2019) Hereditary multiple exostoses: Current insights. *Orthop. Res. Rev.* **11:** 199–211

D'Hulst C & Mérida Á (2010) The priming of storage glucan synthesis from bacteria to plants: current knowledge and new developments. *New Phytol.* **188:** 13–21

Dardelle F, Le Mauff F, Lehner A, Loutelier-Bourhis C, Bardor M, Rihouey C, Causse M, Lerouge P, Driouich A & Mollet J-C (2015) Pollen tube cell walls of wild and domesticated tomatoes contain arabinosylated and fucosylated xyloglucan. *Ann. Bot.* **115:** 55–66

Darriba D, Taboada GL, Doallo R & Posada D (2011) ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27:** 1164–1165

DeAngelis PL (2002) Evolution of glycosaminoglycans and their glycosyltransferases: Implications for the extracellular matrices of animals and the capsules of pathogenic bacteria. *Anat. Rec.* **268:** 317–326

Del-Bem L-E (2018) Xyloglucan evolution and the terrestrialization of green plants. *New Phytol.* **219:** 1150–1153

Deutschmann R & Dekker RFH (2012) From plant biomass to bio-based chemicals: Latest developments in xylan research. *Biotechnol. Adv.* **30:** 1627–1640

Van Dijk ADJ, Bosch D, Ter Braak CJF, Van Der Krol AR & Van Ham RCHJ (2008) Predicting sub-Golgi localization of type II membrane proteins. *Bioinformatics* **24:** 1779–1786

DiRita VJ & Mekalanos JJ (1991) Periplasmic interaction between two membrane regulatory

proteins, ToxR and ToxS, results in signal transduction and transcriptional activation. *Cell* **64:** 29–37

Dobson-Stone C, Cox RD, Lonie L, Southam L, Fraser M, Wise C, Bernier F, Hodgson S, Porter DE, Simpson HR & Monaco AP (2000) Comparison of fluorescent single-strand conformation polymorphism analysis and denaturing high-performance liquid chromatography for detection of EXT1 and EXT2 mutations in hereditary multiple exostoses. *Eur. J. Hum. Genet.* **8:** 24–32

Dobzhansky T (1973) Nothing in Biology Makes Sense except in the Light of Evolution. *Am. Biol. Teach.* **35:** 125–129

Domozych DS, Wells B & Shaw PJ (1991) Basket scales of the green alga, Mesostigma viride: chemistry and ultrastructure. *J. Cell Sci.* **100:** 397–407

Drozdetskiy A, Cole C, Procter J & Barton GJ (2015) JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.* **43:** W389–W394

Eddy SR (2008) A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLoS Comput. Biol.* **4:** e1000069

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Informatics* **23:** 205–211

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7:** e1002195

Edgar RC (2004a) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797

Edgar RC (2004b) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113

Edvardsson E, Singh SK, Yun M-S, Mansfeld A, Hauser M-T & Marchant A (2011) The Plant Glycosyltransferase Family GT64: In Search of a Function. In *Annual Plant Reviews, Volume 41: Plant Polysaccharides*, Ulvskov P (ed) pp 285–303. Chichester: Wiley-Blackwell

Ekstrom A, Taujale R, Mcginn N & Yin Y (2014) PlantCAZyme: a database for plant carbohydrate-active enzymes. *Database* **2014:** 79

El-Battari A, Prorok M, Angata K, Mathieu S, Zerfaoui M, Ong E, Suzuki M, Lombardo D & Fukuda M (2003) Different glycosyltransferases are differentially processed for secretion, dimerization, and autoglycosylation. *Glycobiology* **13:** 941–953

Eliasson A-C (2017) Carbohydrates in Food Boca Raton: CRC Press/Taylor & Francis

Eme L, Sharpe SC, Brown MW & Roger AJ (2014) On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. *Cold Spring Harb. Perspect. Biol.* **6:**

a016139

Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66:** 486–501

Engel BD, Schaffer M, Albert S, Asano S, Plitzko JM & Baumeister W (2015) In situ structural analysis of Golgi intracisternal protein arrays. *Proc. Natl. Acad. Sci. U. S. A.* **112:** 11264–11269

Engler C, Gruetzner R, Kandzia R & Marillonnet S (2009) Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PLoS One* **4:** e5553

Engler C, Youles M, Gruetzner R, Ehnert TM, Werner S, Jones JDG, Patron NJ & Marillonnet S (2014) A Golden Gate modular cloning toolbox for plants. *ACS Synth. Biol.* **3:** 839–843

Esko JD, Kimata K & Lindahl U (2009) Proteoglycans and Sulfated Glycosaminoglycans. In *Essentials of Glycobiology*, Varki A Cummings R & Esko J (eds) New York: Cold Spring Harbor Laboratory Press

Esko JD & Selleck SB (2002) Order Out of Chaos: Assembly of Ligand Binding Sites in Heparan Sulfate. *Annu. Rev. Biochem.* **71:** 435–471

Esko JD & Zhang L (1996) Influence of core protein sequence on glycosaminoglycan assembly. *Curr. Opin. Struct. Biol.* **6:** 663–670

Evtuguin D V., Tomás JL, Silva AMS & Neto CP (2003) Characterization of an acetylated heteroxylan from Eucalyptus globulus Labill. *Carbohydr. Res.* **338:** 597–604

Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW, Seyama Y, Shaw W, Shimizu T, Spener F, van Meer G, VanNieuwenhze MS, White SH, Witztum JL & Dennis EA (2005) A comprehensive classification system for lipids. *Eur. J. Lipid Sci. Technol.* **107:** 337–364

Fang L, Ishikawa T, Rennie EA, Murawska GM, Lao J, Yan J, Tsai AYL, Baidoo EEK, Xu J, Keasling JD, Demura T, Kawai-Yamada M, Scheller H V. & Mortimera JC (2016) Loss of inositol phosphorylceramide sphingolipid mannosylation induces plant immune responses and reduces cellulose content in arabidopsis. *Plant Cell* **28:** 2991–3004

Fanutti C, Gidley MJ & Reid JSG (1991) A xyloglucan-oligosaccharide-specific α-d-xylosidase or exo-oligoxyloglucan-α-xylohydrolase from germinated nasturtium (Tropaeolum majus L.) seeds. *Planta* **184:** 137–147

Felsenstein J (1989) PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166

Feta A, Do AT, Rentzsch F, Technau U & Kusche-Gullberg M (2009) Molecular analysis of heparan sulfate biosynthetic enzyme machinery and characterization of heparan sulfate

structure in Nematostella vectensis. *Biochem. J.* **419:** 585–593

Filiault DL, Ballerini ES, Mandáková T, Aköz G, Derieg NJ, Schmutz J, Jenkins J, Grimwood J, Shu S, Hayes RD, Hellsten U, Barry K, Yan J, Mihaltcheva S, Karafiátová M, Nizhynska V, Kramer EM, Lysak MA, Hodges SA & Nordborg M (2018) The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *eLife* **7:** e36426

Fischer E (1891) Ueber die Configuration des Traubenzuckers und seiner Isomeren. *Berichte der Dtsch. Chem. Gesellschaft* **24:** 1836–1845

Fondeur-Gelinotte M, Lattard V, Oriol R, Mollicone R, Jacquinet J-C, Mulliert G, Gulberti S, Netter P, Magdalou J, Ouzzine M & Fournel-Gigleux S (2006) Phylogenetic and mutational analyses reveal key residues for UDP-glucuronic acid binding and activity of β1,3-glucuronosyltransferase I (GlcAT-I). *Protein Sci.* **15:** 1667–1678

Franková L & Fry SC (2013) Biochemistry and physiological roles of enzymes that 'cut and paste' plant cell-wall polysaccharides. *J. Exp. Bot.* **64:** 3519–3550

Fransson LÅ, Belting M, Jönsson M, Mani K, Moses J & Oldberg Å (2000) Biosynthesis of decorin and glypican. *Matrix Biol.* **19:** 367–376

Fry SC (1986) Cross-Linking of Matrix Polymers in the Growing Cell Walls of Angiosperms. *Annu. Rev. Plant Physiol.* **37:** 165–186

Fry SC (2011) Cell Wall Polysaccharide Composition and Covalent Crosslinking. In *Annual Plant Reviews, Plant Polysaccharides: Biosynthesis and Bioengineering* pp 1–42. Chichester: Wiley-Blackwell

Fry SC, York WS, Albersheim P, Darvill A, Hayashi T, Joseleau J -P, Kato Y, Lorences EP, Maclachlan GA, McNeil M, Mort AJ, Grant Reid JS, Seitz HU, Selvendran RR, Voragen AGJ & White AR (1993) An unambiguous nomenclature for xyloglucan-derived oligosaccharides. *Physiol. Plant.* **89:** 1–3

Fusco C, Nardella G, Fischetto R, Copetti M, Petracca A, Annunziata F, Augello B, D'Asdia MC, Petrucci S, Mattina T, Rella A, Cassina M, Bengala M, Biagini T, Causio FA, Caldarini C, Brancati F, De Luca A, Guarnieri V, Micale L, et al (2019) Mutational spectrum and clinical signatures in 114 families with hereditary multiple osteochondromas: Insights into molecular properties of selected exostosin variants. *Hum. Mol. Genet.* **28:** 2133–2142

Gabler F, Nam S, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN & Alva V (2020) Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinforma.* **72:** e108

Gawkowska D, Cybulska J & Zdunek A (2018) Structure-related gelling of pectins and linking with other natural compounds: A review. *Polymers (Basel).* **10:** 762

Gentile M, Agolini E, Cocciadiferro D, Ficarella R, Ponzi E, Bellacchio E, Antonucci MF & Novelli A (2019) Novel *exostosin-2* missense variants in a family with autosomal recessive exostosin-2-related syndrome: further evidences on the phenotype. *Clin. Genet.* **95:** 165–171

Geshi N, Harholt J, Sakuragi Y, Jensen JK & Scheller HV (2011) Glycosyltransferases of the GT47 Family. In *Annual Plant Reviews, Plant Polysaccharides: Biosynthesis and Bioengineering* pp 265–283. Chichester: Wiley-Blackwell

Gibbons BJ, Roach PJ & Hurley TD (2002) Crystal structure of the autocatalytic initiator of glycogen biosynthesis, glycogenin. *J. Mol. Biol.* **319:** 463–477

Gleeson PA, Teasdale RD & Burke J (1994) Targeting of proteins to the Golgi apparatus. *Glycoconj. J.* **11:** 381–394

Gloster TM (2014) Advances in understanding glycosyltransferases from a structural perspective. *Curr. Opin. Struct. Biol.* **28:** 131–141

Gorin PAJ & Barreto-Bergter E (1983) The Chemistry of Polysaccharides of Fungi and Lichens. In *The Polysaccharides,* Aspinall G. O. (ed) pp 365–409. Academic Press

Gorshkova TA, Gurjanov OP, Mikshina P V., Ibragimova NN, Mokshina NE, Salnikov V V., Ageeva M V., Amenitskii SI, Chernova TE & Chemikosova SB (2010) Specific type of secondary cell wall formed by plant fibers. *Russ. J. Plant Physiol.* **57:** 328–341

Goto M (2007) Bioscience, Biotechnology, and Biochemistry Protein O-Glycosylation in Fungi: Diverse Structures and Multiple Functions. *Biosci. Biotechnol. Biochem.* **71:** 1415–1427

Goubet F, Jackson P, Deery MJ & Dupree P (2002) Polysaccharide analysis using carbohydrate gel electrophoresis. A method to study plant cell wall polysaccharides and polysaccharide hydrolases. *Anal. Biochem.* **300:** 53–68

Grabenhorst E & Conradt HS (1999) The cytoplasmic, transmembrane, and stem regions of glycosyltransferases specify their in vivo functional sublocalization and stability in the Golgi. *J. Biol. Chem.* **274:** 36107–36116

de Graffenried CL & Bertozzi CR (2004) The roles of enzyme localisation and complex formation in glycan assembly within the Golgi apparatus. *Curr. Opin. Cell Biol.* **16:** 356–363

Grantham NJ, Wurman-Rodrich J, Terrett OM, Lyczakowski JJ, Stott K, Iuga D, Simmons TJ, Durand-Tardif M, Brown SP, Dupree R, Busse-Wicher M & Dupree P (2017) An even

pattern of xylan substitution is critical for interaction with cellulose in plant cell walls. *Nat. Plants* **3:** 859–865

Gronnier J, Germain V, Gouguet P, Cacas JL & Mongrand S (2016) GIPC: Glycosyl inositol phospho ceramides, the major sphingolipids on earth. *Plant Signal. Behav.* **11:** e1152438

Grosberg RK & Strathmann RR (2007) The Evolution of Multicellularity: A Minor Major Transition? *Annu. Rev. Ecol. Evol. Syst.* **38:** 621–654

Guan R, Zhao Y, Zhang H, Fan G, Liu X, Zhou W, Shi C, Wang J, Liu W, Liang X, Fu Y, Ma K, Zhao L, Zhang F, Lu Z, Lee SMY, Xu X, Wang J, Yang H, Fu C, et al (2016) Draft genome of the living fossil Ginkgo biloba. *Gigascience* **5:** 49

Guan R, Zhao Y, Zhang H, Fan G, Liu X, Zhou W, Shi C, Wang J, Liu W, Liang X, Fu Y, Ma K, Zhao L, Zhang F, Lu Z, Lee SMY, Xu X, Wang J, Yang H, Fu C, et al (2019) Updated genome assembly of Ginkgo biloba GigaScience Database.

Günl M, Gille S & Pauly M (2010) OLIgo mass profiling (OLIMP) of extracellular polysaccharides. *J. Vis. Exp.* **40:** e2046

Günl M, Neumetzler L, Kraemer F, de Souza A, Schultink A, Pena M, York WS & Pauly M (2011) AXY8 encodes an α-fucosidase, underscoring the importance of apoplastic metabolism on the fine structure of Arabidopsis cell wall polysaccharides. *Plant Cell* **23:** 4025–4040

Günl M & Pauly M (2011) AXY3 encodes a α-xylosidase that impacts the structure and accessibility of the hemicellulose xyloglucan in Arabidopsis plant cell walls. *Planta* **233:** 707–719

Guo L, Elcioglu NH, Mizumoto S, Wang Z, Noyan B, Albayrak HM, Yamada S, Matsumoto N, Miyake N, Nishimura G & Ikegawa S (2017) Identification of biallelic EXTL3 mutations in a novel type of spondylo-epi-metaphyseal dysplasia. *J. Hum. Genet.* **62:** 797–801

Hanes MS, Moremen KW & Cummings RD (2017) Biochemical characterization of functional domains of the chaperone Cosmc. *PLoS One* **12:** e0180242

Hang HC & Bertozzi CR (2005) The chemistry and biology of mucin-type O-linked glycosylation. *Bioorganic Med. Chem.* **13:** 5021–5034

Hanson SR, Best MD & Wong CH (2004) Sulfatases: Structure, mechanism, biological activity, inhibition, and synthetic utility. *Angew. Chemie - Int. Ed.* **43:** 5736–5763

Harholt J, Jensen JK, Sørensen SO, Orfila C, Pauly M & Scheller HV (2006) ARABINAN DEFICIENT 1 is a putative arabinosyltransferase involved in biosynthesis of pectic arabinan in arabidopsis. *Plant Physiol.* **140:** 49–58

Harholt J, Jensen JK, Verhertbruggen Y, Søgaard C, Bernard S, Nafisi M, Poulsen CP, Geshi N, Sakuragi Y, Driouich A, Knox JP & Scheller HV (2012) ARAD proteins associated with pectic Arabinan biosynthesis form complexes when transiently overexpressed in planta. *Planta* **236:** 115–128

Harholt J, Moestrup Ø & Ulvskov P (2016) Why Plants Were Terrestrial from the Beginning. *Trends Plant Sci.* **21:** 96–101

Harrus D, Kellokumpu S & Glumoff T (2018) Crystal structures of eukaryote glycosyltransferases reveal biologically relevant enzyme homooligomers. *Cell. Mol. Life Sci.* **75:** 833–848

Hashimoto K, Madej T, Bryant SH & Panchenko AR (2010) Functional States of Homooligomers: Insights from the Evolution of Glycosyltransferases. *J. Mol. Biol.* **399:** 196–206

Hassinen A, Khoder-Agha F, Khosrowabadi E, Mennerich D, Harrus D, Noel M, Dimova EY, Glumoff T, Harduin-Lepers A, Kietzmann T & Kellokumpu S (2019) A Golgi-associated redox switch regulates catalytic activation and cooperative functioning of ST6Gal-I with B4GalT-I. *Redox Biol.* **24:** 101182

Hatfield RD, Rancour DM & Marita JM (2017) Grass cell walls: A story of cross-linking. *Front. Plant Sci.* **7:** 2056

Held MA, Jiang N, Basu D, Showalter AM & Faik A (2015) Plant Cell Wall Polysaccharides: Structure and Biosynthesis. In *Polysaccharides: Bioactivity and Biotechnology* pp 1–2241. Springer

Hemsley PA, Weimar T, Lilley KS, Dupree P & Grierson CS (2013) A proteomic approach identifies many novel palmitoylated proteins in Arabidopsis. *New Phytol.* **197:** 805–814

Herburger K, Ryan LM, Popper ZA & Holzinger A (2018) Localisation and substrate specificities of transglycanases in charophyte algae relate to development and morphology. *J. Cell Sci.* **131:** jcs203208

Hilz H, de Jong LE, Kabel MA, Verhoef R, Schols HA & Voragen AGJ (2007) Bilberry xyloglucan-novel building blocks containing β-xylose within a complex structure. *Carbohydr. Res.* **342:** 170–181

Hirata T, Mishra SK, Nakamura S, Saito K, Motooka D, Takada Y, Kanzawa N, Murakami Y, Maeda Y, Fujita M, Yamaguchi Y & Kinoshita T (2018) Identification of a Golgi GPI-N-acetylgalactosamine transferase with tandem transmembrane regions in the catalytic domain. *Nat. Commun.* **9:** 405

Hoang DT, Chernomor O, von Haeseler A, Minh BQ & Vinh LS (2018) UFBoot2: Improving

the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35:** 518–522

Hoffman M, Jia Z, Peña MJ, Cash M, Harper A, Blackburn AR, Darvill A & York WS (2005) Structural analysis of xyloglucans in the primary cell walls of plants in the subclass Asteridae. *Carbohydr. Res.* **340:** 1826–1840

Hoh SW, Burnley T & Cowtan K (2020) Current approaches for automated model building into cryo-EM maps using Buccaneer with CCP-EM. *Acta Crystallogr. Sect. D Struct. Biol.* **76:** 531–541

Hol W, Duijnen P Van & Berendsen H (1978) The alpha-helix dipole and the properties of proteins. Cited by me. *Nature* **273:** 443–446

Holm L (2020) DALI and the persistence of protein shape. *Protein Sci.* **29:** 128–140

Honke K (2013) Biosynthesis and biological function of sulfoglycolipids. *Proc. Japan Acad. Ser. B Phys. Biol. Sci.* **89:** 129–138

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N, Moriyama T, Ikeuchi M, Watanabe M, Wada H, Kobayashi K, Saito M, Masuda T, Sasaki-Sekimoto Y, Mashiguchi K, Awai K, et al (2014) Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* **5:** 3978

Hotchkiss AT, Nuñez A, Strahan GD, Chau HK, White AK, Marais JPJ, Hom K, Vakkalanka MS, Di R, Yam KL & Khoo C (2015) Cranberry Xyloglucan Structure and Inhibition of Escherichia coli Adhesion to Epithelial Cells. *J. Agric. Food Chem.* **63:** 5622–5633

Hsieh YSY & Harris PJ (2019) Xylans of red and green algae: What is known about their structures and how they are synthesised? *Polymers (Basel).* **11:** 354

Hu L, Li L, Xie H, Gu Y & Peng T (2011) The Golgi Localization of GOLPH2 (GP73/GOLM1) Is Determined by the Transmembrane and Cytoplamic Sequences. *PLoS One* **6:** e28207

Hu Y & Walker S (2002) Remarkable structural similarities between diverse glycosyltransferases. *Chem. Biol.* **9:** 1287–1296

Iijima K & Hashizume M (2015) Application of Polysaccharides as Structural Materials. *Trends Glycosci. Glycotechnol.* **27:** 67–79

Im H, Sambrook J & Russell DW (2011) The Inoue Method for Preparation and Transformation of Competent E. coli: 'Ultra Competent' Cells. *Bio-101* **1:** e143

Inamori K, Yoshida-Moriguchi T, Hara Y, Anderson ME, Yu L & Campbell KP (2012) Dystroglycan function requires xylosyl- and glucuronyltransferase activities of LARGE. *Science* **335:** 93–96

Ishikawa T, Fang L, Rennie EA, Sechet J, Yan J, Jing B, Moore W, Cahoon EB, Scheller H

V., Kawai-Yamada M & Mortimer JC (2018) GLUCOSAMINE INOSITOLPHOSPHORYLCERAMIDTRANSFERASE1 (GINT1) is a GlcNAc-containing glycosylinositol phosphorylceramide glycosyltransferase. *Plant Physiol.* **177:** 938–952

Ishimaru D, Gotoh M, Takayama S, Kosaki R, Matsumoto Y, Narimatsu H, Sato T, Kimata K, Akiyama H, Shimizu K & Matsumoto K (2016) Large-scale mutational analysis in the EXT1 and EXT2 genes for Japanese patients with multiple osteochondromas. *BMC Genet.* **17:** 52

Izuno A, Hatakeyama M, Nishiyama T, Tamaki I, Shimizu-Inatsugi R, Sasaki R, Shimizu KK & Isagi Y (2016) Genome sequencing of Metrosideros polymorpha (Myrtaceae), a dominant species in various habitats in the Hawaiian Islands with remarkable phenotypic variations. *J. Plant Res.* **129:** 727–736

Jackson C, Clayden S & Reyes-Prieto A (2015) The Glaucophyta: The blue-green plants in a nutshell. *Acta Soc. Bot. Pol.* **84:** 149–165

Jaeken J & Péanne R (2017) What is new in CDG? *J. Inherit. Metab. Dis.* **40:** 569–586

Jaiman A & Thattai M (2020) Golgi compartments enable controlled biomolecular assembly using promiscuous enzymes. *eLife* **9:** e49573

Jarrell KF, Ding Y, Meyer BH, Albers S-V, Kaminski L & Eichler J (2014) N-Linked Glycosylation in Archaea: a Structural, Functional, and Genetic Analysis. *Microbiol. Mol. Biol. Rev.* **78:** 304–341

Jarvis MC (2018) Structure of native cellulose microfibrils, the starting point for nanocellulose manufacture. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **376:** 20170045

Jensen JK, Busse-Wicher M, Poulsen CP, Fangel JU, Smith PJ, Yang J-YY, Peña M-JJ, Dinesen MH, Martens HJ, Melkonian M, Wong GK-SS, Moremen KW, Wilkerson CG, Scheller HV, Dupree P, Ulvskov P, Urbanowicz BR & Harholt J (2018) Identification of an algal xylan synthase indicates that there is functional orthology between algal and plant cell wall biosynthesis. *New Phytol.* **218:** 1049–1060

Jensen JK, Johnson NR & Wilkerson CG (2014) Arabidopsis thaliana IRX10 and two related proteins from psyllium and Physcomitrella patens are xylan xylosyltransferases. *Plant J.* **80:** 207–215

Jensen JK, Schultink A, Keegstra K, Wilkerson CG & Pauly M (2012) RNA-Seq Analysis of Developing Nasturtium Seeds (Tropaeolum majus): Identification and Characterization of an additional Galactosyltransferase Involved in Xyloglucan Biosynthesis. *Mol. Plant* **5:** 984–992

References

Jensen JK, Sørensen SO, Harholt J, Geshi N, Sakuragi Y, Møller I, Zandleven J, Bernal AJ, Jensen NB, Sørensen C, Pauly M, Beldman G, Willats WGT & Scheller HV (2008) Identification of a xylogalacturonan xylosyltransferase involved in pectin biosynthesis in Arabidopsis. *Plant Cell* **20:** 1289–1302

Jia Z, Qin Q, Darvill AG & York WS (2003) Structure of the xyloglucan produced by suspension-cultured tomato cells. *Carbohydr. Res.* **338:** 1197–1208

Jiang N, Wiemels RE, Soya A, Whitley R, Held M & Faik A (2016) Composition, Assembly, and Trafficking of a Wheat Xylan Synthase Complex. *Plant Physiol.* **170:** 1999–2023

Jiao G, Yu G, Zhang J & Ewart H (2011) Chemical Structures and Bioactivities of Sulfated Polysaccharides from Marine Algae. *Mar. Drugs* **9:** 196–223

Johnson RM, Hecht K & Deber CM (2007) Aromatic and cation-π interactions enhance helix-helix association in a membrane environment. *Biochemistry* **46:** 9208–9214

Jones P & Vogt T (2001) Glycosyltransferases in secondary plant metabolism: Tranquilizers and stimulant controllers. *Planta* **213:** 164–174

Ju T & Cummings RD (2014) Core 1 b3Galactosyltransferase (C1GalT1, T-Synthase) and Its Specific Molecular Chaperone Cosmc (C1GalT1C1). In *Handbook of Glycosyltransferases and Related Genes*, Taniguchi N Honke K Fukuda M Narimatsu H Yamaguchi Y & Angata T (eds) pp 149–169. Tokyo: Springer

Kakuda S, Shiba T, Ishiguro M, Tagawa H, Oka S, Kajihara Y, Kawasaki T, Wakatsuki S & Kato R (2004) Structural basis for acceptor substrate recognition of a human glucuronyltransferase, GlcAT-P, an enzyme critical in the biosynthesis of the carbohydrate epitope HNK-1. *J. Biol. Chem.* **279:** 22693–22703

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A & Jermiin LS (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14:** 587–589

Kamath VP, Seto NOL, Compston CA, Hindsgaul O & Palcic MM (1999) Synthesis of the acceptor analog αFuc(1→2)αGal-O(CH2)7 CH3: A probe for the kinetic mechanism of recombinant human blood group B glycosyltransferase. *Glycoconj. J.* **16:** 599–606

Kamide K (2005) Cellulose and cellulose derivatives : molecular characterization and its applications Amsterdam: Elsevier

Kapitonov D & Yu RK (1999) Conserved domains of glycosyltransferases. *Glycobiology* **9:** 961–978

Katayama T, Sakuma A, Kimura T, Makimura Y, Hiratake J, Sakata K, Yamanoi T, Kumagai H & Yamamoto K (2004) Molecular cloning and characterization of Bifidobacterium

bifidum 1,2-α-L-fucosidase (AfcA), a novel inverting glycosidase (glycoside hydrolase family 95). *J. Bacteriol.* **186:** 4885–4893

Katoh K & Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30:** 772–780

Kattke MD, Gosschalk JE, Martinez OE, Kumar G, Gale RT, Cascio D, Sawaya MR, Philips M, Brown ED & Clubb RT (2019) Structure and mechanism of TagA, a novel membrane-associated glycosyltransferase that produces wall teichoic acids in pathogenic bacteria. *PLOS Pathog.* **15:** e1007723

Kattla JJ, Struwe WB, Doherty M, Adamczyk B, Saldova R, Rudd PM & Campbell MP (2011) Protein glycosylation. In *Comprehensive Biotechnology* pp 501–520. Elsevier

Keeling PJ (2013) The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. *Annu. Rev. Plant Biol.* **64:** 583–607

Keeling PJ & Burki F (2019) Progress towards the Tree of Eukaryotes. *Curr. Biol.* **29:** R808–R817

Kellokumpu S, Hassinen A & Glumoff T (2016) Glycosyltransferase complexes in eukaryotes: Long-known, prevalent but still unrecognized. *Cell. Mol. Life Sci.* **73:** 305–325

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H & Phillips DC (1958) A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181:** 662–666

Keppler BD & Showalter AM (2010) IRX14 and IRX14-LIKE, two glycosyl transferases involved in glucuronoxylan biosynthesis and drought tolerance in arabidopsis. *Mol. Plant* **3:** 834–841

El Khadem HS (2012) Carbohydrate Chemistry: Monosaccharides and Their Oligomers San Diego: Academic Press

Khoder-Agha F, Harrus D, Brysbaert G, Lensink MF, Harduin-Lepers A, Glumoff T & Kellokumpu S (2019a) Assembly of B4GALT1/ST6GAL1 heteromers in the Golgi membranes involves lateral interactions via highly charged surface domains. *J. Biol. Chem.* **294:** 14383–14393

Khoder-Agha F, Sosicka P, Conde ME, Hassinen A, Glumoff T, Olczak M & Kellokumpu · Sakari (2019b) N-acetylglucosaminyltransferases and nucleotide sugar transporters form multi-enzyme-multi-transporter assemblies in golgi membranes in vivo. *Cell. Mol. Life Sci.* **76:** 1821–1832

Kim B-T, Kitagawa H, Tanaka J, Tamura J & Sugahara K (2003) In vitro heparan sulfate polymerization: crucial roles of core protein moieties of primer substrates in addition to

the EXT1-EXT2 interaction. *J. Biol. Chem.* **278:** 41618–41623

Kim B-TT, Kitagawa H, Tamura J -i., Saito T, Kusche-Gullberg M, Lindahl U & Sugahara K (2001) Human tumor suppressor EXT gene family members EXTL1 and EXTL3 encode 1,4- N-acetylglucosaminyltransferases that likely are involved in heparan sulfate/ heparin biosynthesis. *Proc. Natl. Acad. Sci.* **98:** 7176–7181

Kim S, Jeon T-J, Oberai A, Yang D, Schmidt JJ & Bowie JU (2005) Transmembrane glycine zippers: Physiological and pathological roles in membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 14278–14283

Kim SJ, Chandrasekar B, Rea AC, Danhof L, Zemelis-Durfee S, Thrower N, Shepard ZS, Pauly M, Brandizzi F & Keegstra K (2020) The synthesis of xyloglucan, an abundant plant cell wall polysaccharide, requires CSLC function. *Proc. Natl. Acad. Sci. U. S. A.* **117:** 20316–20324

Kimanius D, Forsberg BO, Scheres SHW & Lindahl E (2016) Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5:** e18722

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, et al (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451:** 783–788

Kinoshita T, Ohishi K & Takeda J (1997) GPI-Anchor synthesis in mammalian cells: Genes, their products, and a deficiency. *J. Biochem.* **122:** 251–257

Kitagawa H (2019) Unexpected Roles of Exostosin-like 2, EXTL2, in Glycosaminoglycan Biosynthesis and Function. *Trends Glycosci. Glycotechnol.* **31:** SE15–SE17

Kitagawa H & Nadanaka S (2002) Exostoses (Multiple)-Like 1-3 (EXTL1-3). In *Handbook of Glycosyltransferases and Related Genes*, Taniguchi N Honke K Fukuda M Narimatsu H Yamaguchi Y & Angata T (eds) pp 885–903. Tokyo: Springer

Kitagawa H, Shimakawa H & Sugahara K (1999) The tumor suppressor EXT-like gene EXTL2 encodes an α1, 4-N- acetylhexosaminyltransferase that transfers N-acetylgalactosamine and N- acetylglucosamine to the common glycosaminoglycan-protein linkage region: The key enzyme for the chain initiation of hepa. *J. Biol. Chem.* **274:** 13933–13937

Kitagawa H, Uyama T & Sugahara K (2001) Molecular Cloning and Expression of a Human Chondroitin Synthase. *J. Biol. Chem.* **276:** 38721–38726

Klute MJ, Melaṇon P & Dacks JB (2011) Evolution and diversity of the Golgi. *Cold Spring Harb. Perspect. Biol.* **3:** a007849

Koike T, Izumikawa T, Sato B & Kitagawa H (2014) Identification of phosphatase that

dephosphorylates xylose in the glycosaminoglycan-protein linkage region of proteoglycans. *J. Biol. Chem.* **289:** 6695–6708

Koike T, Izumikawa T, Tamura JI & Kitagawa H (2009) FAM20B is a kinase that phosphorylates xylose in the glycosaminoglycan-protein linkage region. *Biochem. J.* **421:** 157–162

Kolmar H, Hennecke F, Götze K, Janzer B, Vogt B, Mayer F & Fritz HJ (1995) Membrane insertion of the bacterial signal transduction protein ToxR and requirements of transcription activation studied by modular replacement of different protein substructures. *EMBO J.* **14:** 3895–3904

Kolset SO, Prydz K & Pejler G (2004) Intracellular proteoglycans. *Biochem. J.* **379:** 217–227

Kolter T (2012) Ganglioside Biochemistry. *ISRN Biochem.* **2012:** 1–36

Kong Y, Peña MJ, Renna L, Avci U, Pattathil S, Tuomivaara ST, Li X, Reiter WD, Brandizzi F, Hahn MG, Darvill AG, York WS & O'neill MA (2015) Galactose-depleted xyloglucan is dysfunctional and leads to dwarfism in arabidopsis. *Plant Physiol.* **167:** 1296–1306

Konishi T, Mitsuishi Y & Kato Y (1998) Analysis of the Oligosaccharide Units of Xyloglucans by Digestion with Isoprimeverose-producing Oligoxyloglucan Hydrolase Followed by Anion-exchange Chromatography. *Biosci. Biotechnol. Biochem.* **62:** 2421–2424

Kreisman LSC & Cobb BA (2012) Infection, inflammation and host carbohydrates: A Glyco-Evasion Hypothesis. *Glycobiology* **22:** 1019–1030

Kreuger J & Kjellén L (2012) Heparan Sulfate Biosynthesis: Regulation and Variability. *J. Histochem. Cytochem.* **60:** 898–907

Krogh A, Rn Larsson BÈ, Von Heijne G & Sonnhammer ELL (2001) Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **305:** 567–580

Kumar M, Carr P & Turner S (2020) An atlas of Arabidopsis protein S-Acylation reveals its widespread role in plant cell organisation of and function. *bioRxiv***:** DOI: 10.1101/2020.05.12.090415

Kuwabara N, Manya H, Yamada T, Tateno H, Kanagawa M, Kobayashi K, Akasaka-Manya K, Hirose Y, Mizuno M, Ikeguchi M, Toda T, Hirabayashi J, Senda T, Endo T & Kato R (2016) Carbohydrate-binding domain of the POMGnT1 stem region modulates O-mannosylation sites of α-dystroglycan. *Proc. Natl. Acad. Sci.* **113:** 9280–9285

Lairson LL, Henrissat B, Davies GJ & Withers SG (2008) Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* **77:** 521–555

Lampugnani ER, Ho YY, Moller IE, Koh PL, Golz JF, Bacic A & Newbigin E (2016) A

glycosyltransferase from Nicotiana alata pollen mediates synthesis of a linear (1,5)-α-L-arabinan when expressed in Arabidopsis. *Plant Physiol.* **170:** 1962–1974

Lao J, Oikawa A, Bromley JR, McInerney P, Suttangkakul A, Smith-Moritz AM, Plahar H, Chiu T-Y, González Fernández-Niño SM, Ebert B, Yang F, Christiansen KM, Hansen SF, Stonebloom S, Adams PD, Ronald PC, Hillson NJ, Hadi MZ, Vega-Sánchez ME, Loqué D, et al (2014) The plant glycosyltransferase clone collection for functional genomics. *Plant J.* **79:** 517–529

Larsbrink J, Izumi A, Ibatullin FM, Nakhai A, Gilbert HJ, Davies GJ & Brumer H (2011) Structural and enzymatic characterization of a glycoside hydrolase family 31 α-xylosidase from Cellvibrio japonicus involved in xyloglucan saccharification. *Biochem. J.* **436:** 567–580

Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, Klinter S, Pudlo NA, Urs K, Koropatkin NM, Creagh AL, Haynes CA, Kelly AG, Cederholm SN, Davies GJ, Martens EC & Brumer H (2014a) A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506:** 498–502

Larsbrink J, Thompson AJ, Lundqvist M, Gardner JG, Davies GJ & Brumer H (2014b) A complex gene locus enables xyloglucan utilization in the model saprophyte Cellvibrio japonicus. *Mol. Microbiol.* **94:** 418–433

Lauc G, Krištić J & Zoldoš V (2014) Glycans - the third revolution in evolution. *Front. Genet.* **5:** 145

Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, Andrade SCS, Sterrer W, Sørensen M V & Giribet G (2019) Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B* **286:** 20190831

Laursen T, Stonebloom SH, Pidatala VR, Birdseye DS, Clausen MH, Mortimer JC & Scheller HV (2018) Bifunctional glycosyltransferases catalyze both extension and termination of pectic galactan oligosaccharides. *Plant J.* **94:** 340–351

Lee C, Teng Q, Huang W, Zhong R & Ye Z-H (2009) The F8H Glycosyltransferase is a Functional Paralog of FRA8 Involved in Glucuronoxylan Biosynthesis in Arabidopsis. *Plant Cell Physiol.* **50:** 812–827

Lee C, Teng Q, Zhong R & Ye ZH (2011) Molecular dissection of Xylan biosynthesis during wood formation in poplar. *Mol. Plant* **4:** 730–747

Lee C, Zhong R, Richardson EA, Himmelsbach DS, McPhail BT & Ye Z-H (2007) The PARVUS Gene is Expressed in Cells Undergoing Secondary Wall Thickening and is Essential for Glucuronoxylan Biosynthesis. *Plant Cell Physiol.* **48:** 1659–1672

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, Porsch M, Quint M, Rensing SA, Soltis DE, Soltis PS, Stevenson DW, Ullrich KK, Wickett NJ, DeGironimo L, Edger PP, et al (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574:** 679–685

Leipelt M, Warnecke D, Zähringer U, Ott C, Müller F, Hube B & Heinz E (2001) Glucosylceramide Synthases, a Gene Family Responsible for the Biosynthesis of Glucosphingolipids in Animals, Plants, and Fungi. *J. Biol. Chem.* **276:** 33621–33629

Lemmon MA, Flanagan JM, Treutlein HR, Zhang J & Engelman DM (1992) Sequence Specificity in the Dimerization of Transmembrane β-Helixes. *Biochemistry* **31:** 12719–12725

Li E, Wimley WC & Hristova K (2012) Transmembrane helix dimerization: Beyond the search for sequence motifs. *Biochim. Biophys. Acta* **1818:** 183–193

Li FW, Brouwer P, Carretero-Paulet L, Cheng S, De Vries J, Delaux PM, Eily A, Koppers N, Kuo LY, Li Z, Simenc M, Small I, Wafula E, Angarita S, Barker MS, Bräutigam A, Depamphilis C, Gould S, Hosmani PS, Huang YM, et al (2018) Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4:** 460–472

Li FW, Nishiyama T, Waller M, Frangedakis E, Keller J, Li Z, Fernandez-Pozo N, Barker MS, Bennett T, Blázquez MA, Cheng S, Cuming AC, de Vries J, de Vries S, Delaux PM, Diop IS, Harrison CJ, Hauser D, Hernández-García J, Kirbis A, et al (2020) Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat. Plants* **6:** 259–272

Li X, Cordero I, Caplan J, Mølhøj M & Reiter WD (2004) Molecular analysis of 10 coding regions from arabidopsis that are homologous to the MUR3 xyloglucan galactosyltransferase. *Plant Physiol.* **134:** 940–950

Liang Z, Geng Y, Ji C, Du H, Wong CE, Zhang Q, Zhang Y, Zhang P, Riaz A, Chachar S, Ding Y, Wen J, Wu Y, Wang M, Zheng H, Wu Y, Demko V, Shen L, Han X, Zhang P, et al (2020) *Mesostigma viride* Genome and Transcriptome Provide Insights into the Origin and Evolution of Streptophyta. *Adv. Sci.* **7:** 1901850

Liepman AH & Cavalier DM (2012) The CELLULOSE SYNTHASE-LIKE A and CELLULOSE SYNTHASE-LIKE C families: Recent advances and future perspectives. *Front. Plant Sci.* **3:** 109

Lira-Navarrete E, Valero-González J, Villanueva R, Martínez-Júlvez M, Tejero T, Merino P, Panjikar S & Hurtado-Guerrero R (2011) Structural Insights into the Mechanism of

Protein O-Fucosylation. *PLoS One* **6:** e25365

Litwack G (2018) Glycogen and Glycogenolysis. In *Human Biochemistry* pp 161–181. Elsevier

Liu J & Mushegian A (2003) Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci.* **12:** 1418–1431

Liu L, Doray B & Kornfeld S (2018) Recycling of Golgi glycosyltransferases requires direct binding to coatomer. *Proc. Natl. Acad. Sci. U. S. A.* **115:** 8984–8989

Liu L, Paulitz J & Pauly M (2015) The presence of fucogalactoxyloglucan and its synthesis in rice indicates conserved functional importance in plants. *Plant Physiol.* **168:** 549–556

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM & Henrissat B (2013) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42:** D490–D495

Lopes FJF, Pauly M, Brommonshenkel SH, Lau EY, Diola V, Passos JL & Loureiro ME (2010) The EgMUR3 xyloglucan galactosyltransferase from Eucalyptus grandis complements the mur3 cell wall phenotype in Arabidopsis thaliana. *Tree Genet. Genomes* **6:** 745–756

Lovegrove A, Edwards CH, De Noni I, Patel H, El SN, Grassby T, Zielke C, Ulmius M, Nilsson L, Butterworth PJ, Ellis PR & Shewry PR (2017) Role of polysaccharides in food, digestion, and health Role of polysaccharides in food, digestion, and health. *Crit. Rev. Food Sci. Nutr.* **57:** 237–253

Lovegrove A, Wilkinson MD, Freeman J, Pellny TK, Tosi P, Saulnier L, Shewry PR & Mitchell RAC (2013) RNA interference suppression of genes in glycosyl transferase families 43 and 47 in wheat starchy endosperm causes large decreases in arabinoxylan content. *Plant Physiol.* **163:** 95–107

Lund CH, Bromley JR, Stenbæk A, Rasmussen RE, Scheller H V. & Sakuragi Y (2015) A reversible Renilla luciferase protein complementation assay for rapid identification of protein–protein interactions reveals the existence of an interaction network involved in xyloglucan biosynthesis in the plant Golgi apparatus. *J. Exp. Bot.* **66:** 85–97

Luo X, Li H, Wu Z, Yao W, Zhao P, Cao D, Yu H, Li K, Poudel K, Zhao D, Zhang F, Xia X, Chen L, Wang Q, Jing D & Cao S (2020) The pomegranate (Punica granatum L.) draft genome dissects genetic divergence between soft- and hard-seeded cultivars. *Plant Biotechnol. J.* **18:** 955–968

Ma Q & Gao X (2019) Categories and biomanufacturing methods of glucosamine. *Appl. Microbiol. Biotechnol.* **103:** 7883–7889

MacKenzie KR, Prestegard JH & Engelman DM (1997) Transmembrane helix dimer: Structure

and implications. *Science* **276:** 131–133

Madson M, Dunand C, Li X, Verma R, Vanzin GF, Caplan J, Shoue DA, Carpita NC & Reiter WD (2003) The MUR3 gene of Arabidopsis encodes a xyloglucan galactosyltransferase that is evolutionarily related to animal exostosins. *Plant Cell* **15:** 1662–1670

Maeda Y, Fujita M & Kinoshita T (2010) GPI-Anchor: Update for Biosynthesis and Remodeling. *Trends Glycosci. Glycotechnol.* **22:** 182–193

Magallon S & Hilu KW (2009) Land plants (Embryophyta). In *The Timetree of Life*, Hedges SB & Kumar S (eds) pp 133–137. Oxford

Malgas S, Mafa MS, Mkabayi L & Pletschke BI (2019) A mini review of xylanolytic enzymes with regards to their synergistic interactions during hetero-xylan degradation. *World J. Microbiol. Biotechnol.* **35:** 187

Mani K & Logan DT (2018) Email to Louis Wilson.

Marinas M, Sa E, Rojas MM, Moalem M, Urbano FJ, Guillou C & Rallo L (2010) A nuclear magnetic resonance (1 H and 13 C) and isotope ratio mass spectrometry (d 13 C , d 2 H and d 18 O) study of Andalusian olive oils. *Rapid Commun. Mass Spectrom.* **24:** 1457–1466

Martinez-Fleites C, Proctor M, Roberts S, Bolam DN, Gilbert HJ & Davies GJ (2006) Insights into the Synthesis of Lipopolysaccharide and Antibiotics through the Structures of Two Retaining Glycosyltransferases from Family GT4. *Chem. Biol.* **13:** 1143–1152

McCaughey J & Stephens DJ (2019) ER-to-Golgi Transport: A Sizeable Problem. *Trends Cell Biol.* **29:** 940–953

McCormick C, Duncan G, Goutsos KT & Tufaro F (2000) The putative tumor suppressors EXT1 and EXT2 form a stable complex that accumulates in the Golgi apparatus and catalyzes the synthesis of heparan sulfate. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 668–673

Meents MJ, Motani S, Mansfield SD & Lacey Samuels A (2019) Organization of xylan production in the Golgi during secondary cell wall biosynthesis. *Plant Physiol.* **181:** 527–546

Mendoza F, Gómez H, Lluch JM & Masgrau L (2016) α1,4-N-Acetylhexosaminyltransferase EXTL2: The Missing Link for Understanding Glycosidic Bond Biosynthesis with Retention of Configuration. *ACS Catal.* **6:** 2577–2589

Mendoza F, Lluch JM & Masgrau L (2017) Computational insights into active site shaping for substrate specificity and reaction regioselectivity in the EXTL2 retaining glycosyltransferase. *Org. Biomol. Chem.* **15:** 9095–9107

Merzendorfer H (2011) The cellular basis of chitin synthesis in fungi and insects: Common

principles and differences. *Eur. J. Cell Biol.* **90:** 759–769

Mikkelsen MD, Harholt J, Ulvskov P, Johansen IE, Fangel JU, Doblin MS, Bacic A & Willats WGT (2014) Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. *Ann. Bot.* **114:** 1217–1236

Miller VL, Taylor RK & Mekalanos JJ (1987) Cholera toxin transcriptional activator ToxR is a transmembrane DNA binding protein. *Cell* **48:** 271–279

Misevic GN & Burger MM (1993) Carbohydrate-carbohydrate interactions of a novel acidic glycan can mediate sponge cell adhesion. *J. Biol. Chem.* **268:** 4922–4929

Mizrachi E, Mansfield SD & Myburg AA (2012) Cellulose factories: advancing bioenergy production from forest trees. *New Phytol.* **194:** 54–62

Møller SR, Yi X, Velásquez SM, Gille S, Hansen PLM, Poulsen CP, Olsen CE, Rejzek M, Parsons H, Zhang Y, Wandall HH, Clausen H, Field RA, Pauly M, Estevez JM, Harholt J, Ulvskov P & Petersen BL (2017) Identification and evolution of a plant cell wall specific glycoprotein glycosyl transferase, ExAD. *Sci. Rep.* **7:** 45341

Moore DT, Berger BW & DeGrado WF (2008) Protein-Protein Interactions in the Membrane: Sequence, Structural, and Biological Motifs. *Structure* **16:** 991–1001

Moracci M, Ponzano BC, Trincone A, Fusco S, De Rosa M, Van Der Oost J, Sensen CW, Charlebois RL & Rossi M (2000) Identification and molecular characterization of the first α-xylosidase from an Archaeon. *J. Biol. Chem.* **275:** 22082–22089

Moremen KW & Haltiwanger RS (2019) Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat. Chem. Biol.* **15:** 853–864

Morise T, Yano Y & Matsuzaki K (2020) Interplay between Amino Acid Sequences and Lipid Compositions in the GXXXG-Mediated Parallel Self-Association of Transmembrane Helices as Revealed by Single-pair Fret. *Biophys. J.* **118:** 368a

Morita I, Kizuka Y, Kakuda S & Oka S (2008) Expression and Function of the HNK-1 Carbohydrate. *J. Biochem.* **143:** 719–724

Mortimer JC, Faria-Blanc N, Yu X, Tryfona T, Sorieul M, Ng YZ, Zhang Z, Stott K, Anders N & Dupree P (2015) An unusual xylan in Arabidopsis primary cell walls is synthesised by GUX3, IRX9L, IRX10L and IRX14. *Plant J.* **83:** 413–426

Moss GP, Smith PAS & Tavernier D (1995) Glossary of class names of organic compounds and reactive intermediates based on structure (IUPAC recommendations 1995). *Pure Appl. Chem.* **67:** 1307–1375

Mourão PA, Vilanova E & Soares PA (2018) Unveiling the structure of sulfated fucose-rich polysaccharides via nuclear magnetic resonance spectroscopy. *Curr. Opin. Struct. Biol.*

50: 33–41

Mueller BK, Subramaniam S & Senes A (2014) A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα–H hydrogen bonds. *Proc. Natl. Acad. Sci.* **111:** E888–E895

Nadanaka S, Purunomo E, Takeda N, Tamura JI & Kitagawa H (2014) Heparan sulfate containing unsubstituted glucosamine residues: Biosynthesis and heparanase-inhibitory activity. *J. Biol. Chem.* **289:** 15231–15243

Nadanaka S, Zhou S, Kagiyama S, Shoji N, Sugahara K, Sugihara K, Asano M & Kitagawa H (2013) EXTL2, a member of the EXT family of tumor suppressors, controls glycosaminoglycan biosynthesis in a xylose kinase-dependent manner. *J. Biol. Chem.* **288:** 9321–9333

Ndeh D & Gilbert HJ (2018) Biochemistry of complex glycan depolymerisation by the human gut microbiota. *FEMS Microbiol. Rev.* **42:** 146–164

Neelamegham S, Aoki-Kinoshita K, Bolton E, Frank M, Lisacek F, Lütteke T, O'boyle N, Packer NH, Stanley P, Toukach P & Varki A (2019) Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **29:** 620–624

Neff EP (2018) What is a lab animal? *Lab Anim. (NY).* **47:** 223–227

Neff MM, Turk E & Kalishman M (2002) Web-based primer design for single nucleotide polymorphism analysis. *Trends Genet.* **18:** 613–615

Nelson BK, Cai X & Nebenführ A (2007) A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. *Plant J.* **51:** 1126–1136

Nguyen L-T, Schmidt HA, von Haeseler A & Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32:** 268–274

Nielsen C (2019) Early animal evolution: a morphologist's view. *R. Soc. Open Sci.* **6:** 190638

Niemann MCE, Weber H, Hluska T, Leonte G, Anderson SM, Novák O, Senes A & Werner T (2018) The cytokinin oxidase/dehydrogenase CKX1 is a membrane-bound protein requiring homooligomerization in the endoplasmic reticulum for its cellular activity. *Plant Physiol.* **176:** 2024–2039

Nilsson T, Slusarewicz P, Hoe MH & Warren G (1993) Kin recognition. *FEBS Lett.* **330:** 1–4

Noborn F, Gomez Toledo A, Green A, Nasir W, Sihlbom C, Nilsson J & Larson G (2016) Site-specific identification of heparan and chondroitin sulfate glycosaminoglycans in hybrid proteoglycans. *Sci. Rep.* **6:** 34537

Novo-Uzal E, Pomar F, Gómez Ros L V., Espiñeira JM & Ros Barceló A (2012) Evolutionary

History of Lignins Academic Press Inc.

Ohashi H, Ohashi T, Misaki R & Fujiyama K (2018) *Arabidopsis thaliana* α1,2-L-fucosyltransferase catalyzes the transfer of L-galactose to xyloglucan oligosaccharides. *FEBS Lett.* **593:** 187–194

Oikawa A, Lund CH, Sakuragi Y & Scheller H V. (2013) Golgi-localized enzyme complexes for plant cell wall biosynthesis. *Trends Plant Sci.* **18:** 49–58

Ori A, Wilkinson MC & Fernig DG (2011) A systems biology approach for the investigation of the heparin/heparan sulfate interactome. *J. Biol. Chem.* **286:** 19892–19904

Osmani SA, Bak S & Møller BL (2009) Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. *Phytochemistry* **70:** 325–347

Oud MM, Tuijnenburg P, Hempel M, van Vlies N, Ren Z, Ferdinandusse S, Jansen MH, Santer R, Johannsen J, Bacchelli C, Alders M, Li R, Davies R, Dupuis L, Cale CM, Wanders RJA, Pals ST, Ocaka L, James C, Müller I, et al (2017) Mutations in EXTL3 Cause Neuro-immuno-skeletal Dysplasia Syndrome. *Am. J. Hum. Genet.* **100:** 281–296

Ouzzine M, Gulberti S, Levoin N, Netter P, Magdalou J & Fournel-Gigleux S (2002) The donor substrate specificity of the human β,3-glucuronosyltransferase I toward UDP-glucuronic acid is determined by two crucial histidine and arginine residues. *J. Biol. Chem.* **277:** 25439–25445

Ouzzine M, Gulberti S, Netter P, Magdalou J & Fournel-Gigleux S (2000) Structure-function of human Gal beta 1,3-glucuronosyltransferase (GlcAT-I) Dimerization and functional activity are mediated by two crucial cysteine Residues. *J. Biol. Chem.* **275:** 28254–28260

Paganini C, Costantini R, Superti-Furga A, Rossi A & Rossi CA (2019) Bone and connective tissue disorders caused by defects in glycosaminoglycan biosynthesis: a panoramic view. *FEBS J.* **286:** 3008–3032

Park YB & Cosgrove DJ (2014) Xyloglucan and its Interactions with Other Components of the Growing Cell Wall. *Plant Cell Physiol.* **56:** 180–194

Parsons HT, Stevens TJ, McFarlane HE, Vidal-Melgosa S, Griss J, Lawrence N, Butler R, Sousa MML, Salemi M, Willats WGT, Petzold CJ, Heazlewood JL & Lilley KS (2019) Separating Golgi proteins from cis to trans reveals underlying properties of cisternal localization. *Plant Cell* **31:** 2010–2034

Patenaude SI, Seto NOL, Borisova SN, Szpacenko A, Marcus SL, Palcic MM & Evans S V. (2002) The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat. Struct. Biol.* **9:** 685–690

Patron NJ, Orzaez D, Marillonnet S, Warzecha H, Matthewman C, Youles M, Raitskin O, Leveau A, Farré G, Rogers C, Smith A, Hibberd J, Webb AAR, Locke J, Schornack S, Ajioka J, Baulcombe DC, Zipfel C, Kamoun S, Jones JDG, et al (2015) Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *New Phytol.* **208:** 13–19

Paulick MG & Bertozzi CR (2008) The glycosylphosphatidylinositol anchor: A complex membrane-anchoring structure for proteins. *Biochemistry* **47:** 6991–7000

Pauly M, Albersheim P, Darvill A & York WS (1999) Molecular domains of the cellulose/xyloglucan network in the cell walls of higher plants. *Plant J.* **20:** 629–639

Pauly M, Gille S, Liu L, Mansoori N, de Souza A, Schultink A & Xiong G (2013) Hemicellulose biosynthesis. *Planta* **238:** 627–642

Pauly M & Keegstra K (2008) Cell-wall carbohydrates and their modification as a resource for biofuels. *Plant J.* **54:** 559–568

Pauly M & Keegstra K (2016) Biosynthesis of the Plant Cell Wall Matrix Polysaccharide Xyloglucan. *Annu. Rev. Plant Biol.* **67:** 235–259

Pedersen LC, Darden TA & Negishi M (2002) Crystal structure of beta 1,3-glucuronyltransferase I in complex with active donor substrate UDP-GlcUA. *J. Biol. Chem.* **277:** 21869–21873

Pedersen LC, Dong J, Taniguchi F, Kitagawa H, Krahn JM, Pedersen LG, Sugahara K & Negishi M (2003) Crystal structure of an α1,4-N-acetylhexosaminyltransferase (EXTL2), a member of the exostosin gene family involved in heparan sulfate biosynthesis. *J. Biol. Chem.* **278:** 14420–14428

Pedersen LC, Tsuchida K, Kitagawa H, Sugahara K, Darden TA & Negishi M (2000) Heparan/chondroitin sulfate biosynthesis. Structure and mechanism of human glucuronyltransferase I. *J. Biol. Chem.* **275:** 34580–34585

Pelaseyed T, Bergström JH, Gustafsson JK, Ermund A, Birchenough GMH, Schütte A, van der Post S, Svensson F, Rodríguez-Piñeiro AM, Nyström EEL, Wising C, Johansson ME V. & Hansson GC (2014) The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system. *Immunol. Rev.* **260:** 8–20

Pellny TK, Patil A, Wood AJ, Freeman J, Halsey K, Plummer A, Kosik O, Temple H, Collins JD, Dupree P, Berry S, Shewry PR, Lovegrove A, Phillips AL & Mitchell RAC (2020) Loss of TaIRX9b gene function in wheat decreases chain length and amount of arabinoxylan in grain but increases cross-linking. *Plant Biotechnol. J.* **18:** 2316–2327

References

Peña MJ, Kong Y, York WS & O'Neill MA (2012) A galacturonic acid-containing xyloglucan is involved in arabidopsis root hair tip growthw. *Plant Cell* **24:** 4511–4524

Peña MJ, Kulkarni AR, Backe J, Boyd M, O'Neill MA & York WS (2016) Structural diversity of xylans in the cell walls of monocots. *Planta* **244:** 589–606

Peña MJ, Zhong R, Zhou G-K, Richardson EA, O'Neill MA, Darvill AG, York WS & Ye Z-H (2007) Arabidopsis irregular xylem8 and irregular xylem9: implications for the complexity of glucuronoxylan biosynthesis. *Plant Cell* **19:** 549–563

Pérez S & Bertoft E (2010) The molecular structures of starch components and their contribution to the architecture of starch granules: A comprehensive review. *Starch - Stärke* **62:** 389–420

Petrik DL, Tryfona T, Dupree P & Anderson CT (2020) BdGT43B2 functions in xylan biosynthesis and is essential for seedling survival in Brachypodium distachyon. *Plant Direct* **4:** 1–16

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC & Ferrin TE (2004) UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25:** 1605–1612

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH & Ferrin TE (2021) UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Sci.* **30:** 70–82

Pogozheva ID & Lomize AL (2018) Evolution and adaptation of single-pass transmembrane proteins. *Biochim. Biophys. Acta - Biomembr.* **1860:** 364–377

Polashock J, Zelzion E, Fajardo D, Zalapa J, Georgi L, Bhattacharya D & Vorsa N (2014) The American cranberry: First insights into the whole genome of a species adapted to bog habitat. *BMC Plant Biol.* **14:** 165

Pomin V (2014) Holothurian Fucosylated Chondroitin Sulfate. *Mar. Drugs* **12:** 232–254

Popper ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B & Stengel DB (2011) Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu. Rev. Plant Biol.* **62:** 567–590

Pothukuchi P, Agliarulo I, Russo D, Rizzo R, Russo F & Parashuraman S (2019) Translation of genome to glycome: role of the Golgi apparatus. *FEBS Lett.* **593:** 2390–2411

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R & Finn RD (2018) HMMER web server: 2018 update. *Nucleic Acids Res.* **46:** W200–W204

Price MN, Dehal PS, Arkin AP, Rojas M & Brodie E (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **5:** e9490

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inze D, Mueller-Roeber B & Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.* **43:** D974–D981

Prydz K (2015) Determinants of Glycosaminoglycan (GAG) Structure. *Biomolecules* **5:** 2003–2022

Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH, Schneider H, Pisani D & Donoghue PCJ (2018) The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Curr. Biol.* **28:** 733–745

Qi X, Kuo LY, Guo C, Li H, Li Z, Qi J, Wang L, Hu Y, Xiang J, Zhang C, Guo J, Huang CH & Ma H (2018) A well-resolved fern nuclear phylogeny reveals the evolution history of numerous transcription factor families. *Mol. Phylogenet. Evol.* **127:** 961–977

Qian R, Chen C & Colley KJ (2001) Location and mechanism of alpha 2,6-sialyltransferase dimer formation. Role of cysteine residues in enzyme dimerization, localization, activity, and processing. *J. Biol. Chem.* **276:** 28641–28649

Rambaut A & Drummond AJ (2018) FigTree v1.4.4.

Ranwez V, Douzery EJP, Cambon C, Chantret N & Delsuc F (2018) MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol. Biol. Evol.* **35:** 2582–2584

Ranwez V, Harispe S, Delsuc F & Douzery EJP (2011) MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PLoS One* **6:** 22594

Rao VSR, Qasba PK, Balaji P V & Chandrasekaran R (1998) Conformation of Carbohydrates CRC Press

Ratke C, Pawar PM-A, Balasubramanian VK, Naumann M, Duncranz ML, Derba-Maceluch M, Gorzsás A, Endo S, Ezcurra I & Mellerowicz EJ (2015) *Populus GT43* family members group into distinct sets required for primary and secondary wall xylan biosynthesis and include useful promoters for wood modification. *Plant Biotechnol. J.* **13:** 26–37

Ratke C, Terebieniec BK, Winestrand S, Derba-Maceluch M, Grahn T, Schiffthaler B, Ulvcrona T, Özparpucu M, Rüggeberg M, Lundqvist S-O, Street NR, Jönsson LJ & Mellerowicz EJ (2018) Downregulating aspen xylan biosynthetic GT43 genes in developing wood stimulates growth via reprograming of the transcriptome. *New Phytol.* **219:** 230–245

Ray B, Loutelier-Bourhis C, Lange C, Condamine E, Driouich A & Lerouge P (2004) Structural investigation of hemicellulosic polysaccharides from Argania spinosa:

References

Characterisation of a novel xyloglucan motif. *Carbohydr. Res.* **339:** 201–208

Reily C, Stewart TJ, Renfrow MB & Novak J (2019) Glycosylation in health and disease. *Nat. Rev. Nephrol.* **15:** 346–366

Ren Y, Hansen SF, Ebert B, Lau J & Scheller HV (2014) Site-Directed Mutagenesis of IRX9, IRX9L and IRX14 Proteins Involved in Xylan Biosynthesis: Glycosyltransferase Activity Is Not Required for IRX9 Function in Arabidopsis. *PLoS One* **9:** e105014

Rennie EA & Scheller HV (2014) Xylan biosynthesis. *Curr. Opin. Biotechnol.* **26:** 100–107

Richmond TA & Somerville CR (2000) The Cellulose Synthase Superfamily. *Plant Physiol.* **124:** 495–498

Robyt JF (1998) Essentials of carbohydrate chemistry New York: Springer Verlag

Rocha J, Cicéron F, de Sanctis D, Lelimousin M, Chazalet V, Lerouxel O & Breton C (2016) Structure of arabidopsis thaliana FUT1 reveals a variant of the GT-B class fold and provides insight into xyloglucan fucosylation. *Plant Cell* **28:** 2352–2364

Roseman S (2001) Reflections on Glycobiology. *J. Biol. Chem.* **276:** 41527–41542

Roth AF, Wan J, Bailey AO, Sun B, Kuchar JA, Green WN, Phinney BS, Yates JR & Davis NG (2006) Global Analysis of Protein Palmitoylation in Yeast. *Cell* **125:** 1003–1013

Roy A, Kucukural A & Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5:** 725–738

Russ WP & Engelman DM (1999) TOXCAT: A measure of transmembrane helix association in a biological membrane. *Proc. Natl. Acad. Sci. U. S. A.* **96:** 863–868

Russ WP & Engelman DM (2000) The GxxxG motif: A framework for transmembrane helix-helix association. *J. Mol. Biol.* **296:** 911–919

Sarkar P, Bosneaga E & Auer M (2009) Plant cell walls throughout evolution: towards a molecular understanding of their design principles. *J. Exp. Bot.* **60:** 3615–3635

Sarrazin S, Lamanna WC & Esko JD (2011) Heparan Sulfate Proteoglycans. *Cold Spring Harb. Perspect. Biol.* **3:** a004952

Schalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS & Cavalier-Smith T (2008) Multigene phylogeny of Choanozoa and the origin of animals. *PLoS One* **3:** e2098

Scheller HV & Ulvskov P (2010) Hemicelluloses. *Annu. Rev. Plant Biol.* **61:** 263–289

Scheres SHW (2012) RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180:** 519–530

Schmitz KR, Liu J, Li S, Setty TG, Wood CS, Burd CG & Ferguson KM (2008) Golgi Localization of Glycosyltransferases Requires a Vps74p Oligomer. *Dev. Cell* **14:** 523–

534

Schnaar RL (2016) Glycobiology simplified: diverse roles of glycan recognition in inflammation. *J. Leukoc. Biol.* **99:** 825–838

Schneider CA, Rasband WS & Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9:** 671–675

Schoberer J, König J, Veit C, Vavra U, Liebminger E, Botchway SW, Altmann F, Kriechbaumer V, Hawes C & Strasser R (2019a) A signal motif retains Arabidopsis ER-α-mannosidase I in the cis-Golgi and prevents enhanced glycoprotein ERAD. *Nat. Commun.* **10:** 3701

Schoberer J, Liebminger E, Botchway SW, Strasser R & Hawes C (2013) Time-resolved fluorescence imaging reveals differential interactions of N-glycan processing enzymes across the Golgi stack in planta. *Plant Physiol.* **161:** 1737–1754

Schoberer J, Liebminger E, Vavra U, Veit C, Grünwald-Gruber C, Altmann F, Botchway SW & Strasser R (2019b) The Golgi localization of GnTI requires a polar amino acid residue within its transmembrane domain. *Plant Physiol.* **180:** 859–873

Schoberer J & Strasser R (2011) Sub-Compartmental Organization of Golgi-Resident N-Glycan Processing Enzymes in Plants. *Mol. Plant* **4:** 220–228

Schrödinger LLC (2020) The PyMOL Molecular Graphics System, Version 2.4.

Schultink A, Cheng K, Park YB, Cosgrove DJ & Pauly M (2013) The identification of two arabinosyltransferases from tomato reveals functional equivalency of xyloglucan side chain substituents. *Plant Physiol.* **163:** 86–94

Schultink A, Liu L, Zhu L & Pauly M (2014) Structural Diversity and Function of Xyloglucan Sidechain Substituents. *Plants* **3:** 526–542

Segrest JP, De Loof H, Dohlman JG, Brouillette CG & Anantharamaiah GM (1990) Amphipathic helix motif: Classes and properties. *Proteins Struct. Funct. Genet.* **8:** 103–117

Senes A, Gerstein M & Engelman DM (2000) Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and association with β-branched residues at neighboring positions. *J. Mol. Biol.* **296:** 921–936

Sethi MK, Buettner FFR, Ashikov A, Krylov VB, Takeuchi H, Nifantiev NE, Haltiwanger RS, Gerardy-Schahn R & Bakker H (2012) Molecular cloning of a xylosyltransferase that transfers the second xylose to O-glucosylated epidermal growth factor repeats of notch. *J. Biol. Chem.* **287:** 2739–2748

Seto NOLL, Compston CA, Evans S V., Bundle DR, Narang SA & Palcic MM (1999) Donor

substrate specificity of recombinant human blood group A, B and hybrid A/B glycosyltransferases expressed in Escherichia coli. *Eur. J. Biochem.* **259:** 770–775

Sharpe HJ, Stevens TJ & Munro S (2010) A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* **142:** 158–169

Shatalov AA, Evtuguin D V. & Pascoal Neto C (1999) (2-O-α-D-Galactopyranosyl-4-O-methyl-α-D-glucurono)-D-xylan from Eucalyptus globulus Labill. *Carbohydr. Res.* **320:** 93–99

Shi YR, Wu JY, Tsai FJ, Lee CC & Tsai CH (2000) An R223P mutation in EXT2 gene causes hereditary multiple exostoses. *Hum. Mutat.* **15:** 390–391

Shiba T, Kakuda S, Ishiguro M, Morita I, Oka S, Kawasaki T, Wakatsuki S & Kato R (2006) Crystal structure of GlcAT-S, a human glucuronyltransferase, involved in the biosynthesis of the HNK-1 carbohydrate epitope. *Proteins Struct. Funct. Bioinforma.* **65:** 499–508

Shimada TL, Shimada T & Hara-Nishimura I (2010) A rapid and non-destructive screenable marker, FAST, for identifying transformed seeds of Arabidopsis thaliana. *Plant J.* **61:** 519–528

Simmons TJ, Frandsen KEH, Ciano L, Tryfona T, Lenfant N, Poulsen JC, Wilson LFL, Tandrup T, Tovborg M, Schnorr K, Johansen KS, Henrissat B, Walton PH, Lo Leggio L & Dupree P (2017) Structural and electronic determinants of lytic polysaccharide monooxygenase reactivity on polysaccharide substrates. *Nat. Commun.* **8:** 1064

Simmons TJ, Mortimer JC, Bernardinelli OD, Pöppler A-C, Brown SP, DeAzevedo ER, Dupree R & Dupree P (2016) Folding of xylan onto cellulose fibrils in plant cell walls revealed by solid-state NMR. *Nat. Commun.* **7:** 13902

Sinnott M (2007) Carbohydrate chemistry and biochemistry: structure and mechanism. Cambridge: RSC Publishing

Smith BG & Harris PJ (1999) The polysaccharide composition of Poales cell walls: Poaceae cell walls are not unique. *Biochem. Syst. Ecol.* **27:** 33–53

Smith PJ, Wang H-T, York WS, Peña MJ & Urbanowicz BR (2017) Designer biomass for next-generation biorefineries: leveraging recent insights into xylan structure and biosynthesis. *Biotechnol. Biofuels* **10:** 286

Sobhany M, Dong J & Negishi M (2005) Two-step mechanism that determines the donor binding specificity of human UDP-N-acetylhexosaminyltransferase. *J. Biol. Chem.* **280:** 23441–23445

Solís D, Bovin N V., Davis AP, Jiménez-Barbero J, Romero A, Roy R, Smetana K & Gabius HJ (2015) A guide into glycosciences: How chemistry, biochemistry and biology

cooperate to crack the sugar code. *Biochim. Biophys. Acta - Gen. Subj.* **1850:** 186–235

Soltis PS, Folk RA & Soltis DE (2019) Darwin review: angiosperm phylogeny and evolutionary radiations. *Proc. R. Soc. B Biol. Sci.* **286:** 20190099

Song L, Zeng W, Wu A, Picard K, Lampugnani ER, Cheetamun R, Beahan C, Cassin A, Lonsdale A, Doblin MS & Bacic A (2015) Asparagus Spears as a Model to Study Heteroxylan Biosynthesis during Secondary Wall Development. *PLoS One* **10:** e0123878

Sørensen I, Domozych D & Willats WGT (2010) How have plant cell walls evolved? *Plant Physiol.* **153:** 366–372

Sørensen I, Pettolino FA, Bacic A, Ralph J, Lu F, O'Neill MA, Fei Z, Rose JKC, Domozych DS & Willats WGT (2011) The charophycean green algae provide insights into the early origins of plant cell walls. *Plant J.* **68:** 201–211

Sorieul M, Dickson A, Hill S & Pearson H (2016) Plant Fibre: Molecular Structure and Biomechanical Properties, of a Complex Living Material, Influencing Its Deconstruction towards a Biobased Composite. *Materials (Basel).* **9:** 618

Soto MJ, Urbanowicz BR & Hahn MG (2019) Plant Fucosyltransferases and the Emerging Biological Importance of Fucosylated Plant Structures. *CRC. Crit. Rev. Plant Sci.* **38:** 327–338

Sousa VL, Brito C, Costa T, Lanoix J, Nilsson T & Costa J (2003) Importance of Cys, Gln, and Tyr from the transmembrane domain of human α3/4 fucosyltransferase III for its localization and sorting in the Golgi of baby hamster kidney cells. *J. Biol. Chem.* **278:** 7624–7629

Soza VL, Lindsley D, Waalkes A, Ramage E, Patwardhan RP, Burton JN, Adey A, Kumar A, Qiu R, Shendure J & Hall B (2019) The Rhododendron Genome and Chromosomal Organization Provide Insight into Shared Whole-Genome Duplications across the Heath Family (Ericaceae). *Genome Biol. Evol.* **11:** 3353–3371

Sparkes IA, Runions J, Kearns A & Hawes C (2006) Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat. Protoc.* **1:** 2019–2025

Sparr E, Ash WL, Nazarov P V., Rijkers DTS, Hemminga MA, Tieleman DP & Killian JA (2005) Self-association of transmembrane α-helices in model membranes: Importance of helix orientation and role of hydrophobic mismatch. *J. Biol. Chem.* **280:** 39324–39331

Stace CA (1992) Plant Taxonomy and Biosystematics Cambridge: Cambridge University Press

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688–2690

References

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313

Stanley P (2011) Golgi Glycosylation. *Cold Spring Harb. Perspect. Biol.* **3:** a005199

Steindorf D & Schneider D (2016) In vivo selection of heterotypically interacting transmembrane helices: Complementary helix surfaces, rather than conserved interaction motifs, drive formation of transmembrane hetero-dimers. *Biochim. Biophys. Acta - Biomembr.* **1859:** 245–256

Stetten D (1960) Biosynthesis and pyrophosphate. *Am. J. Med.* **28:** 867–870

Sukhwal A & Sowdhamini R (2013) Oligomerisation status and evolutionary conservation of interfaces of protein structural domain superfamilies. *Mol. Biosyst.* **9:** 1652–1661

Sun Q, Ju T & Cummings RD (2011) The transmembrane domain of the molecular chaperone cosmc directs its localization to the endoplasmic reticulum. *J. Biol. Chem.* **286:** 11529–11542

Takahara H & Matsuda K (1976) The Structure of Neurospora crassa Glycogen. *Agric. Biol. Chem.* **40:** 1699–1703

Takahashi I, Noguchi N, Nata K, Yamada S, Kaneiwa T, Mizumoto S, Ikeda T, Sugihara K, Asano M, Yoshikawa T, Yamauchi A, Shervani NJ, Uruno A, Kato I, Unno M, Sugahara K, Takasawa S, Okamoto H & Sugawara A (2009) Important role of heparan sulfate in postnatal islet growth and insulin secretion. *Biochem. Biophys. Res. Commun.* **383:** 113–118

Tammi RH, Passi AG, Rilla K, Karousou E, Vigetti D, Makkonen K & Tammi MI (2011) Transcriptional and post-translational regulation of hyaluronan synthesis. *FEBS J.* **278:** 1419–1428

Tamura K, Shimada T, Kondo M, Nishimura M & Hara-Nishimura I (2005) KATAMARI1/MURUS3 is a novel Golgi membrane protein that is required for endomembrane organization in Arabidopsis. *Plant Cell* **17:** 1764–1776

Tan J, Miao Z, Ren C, Yuan R, Tang Y, Zhang X, Han Z & Ma C (2018) Evolution of intron-poor clades and expression patterns of the glycosyltransferase family 47. *Planta* **247:** 745–760

Tan L, Eberhard S, Pattathil S, Warder C, Glushka J, Yuan C, Hao Z, Zhu X, Avci U, Miller JS, Baldwin D, Pham C, Orlando R, Darvill A, Hahn MG, Kieliszewski MJ & Mohnena D (2013) An Arabidopsis cell wall proteoglycan consists of pectin and arabinoxylan covalently linked to an arabinogalactan protein. *Plant Cell* **25:** 270–287

Taujale R, Venkat A, Huang L-C, Zhou Z, Yeung W, Rasheed KM, Li S, Edison AS, Moremen

KW & Kannan N (2020) Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases. *eLife* **9:** e54532

Taujale R & Yin Y (2015) Glycosyltransferase Family 43 Is Also Found in Early Eukaryotes and Has Three Subfamilies in Charophycean Green Algae. *PLoS One* **10:** e0128409

Teese MG & Langosch D (2015) Role of GxxxG Motifs in Transmembrane Domain Interactions. *Biochemistry* **54:** 5125–5135

Telford MJ, Budd GE & Philippe H (2015) Phylogenomic insights into animal evolution. *Curr. Biol.* **25:** R876–R887

Thomas LH, Trevor Forsyth V, Šturcová A, Kennedy CJ, May RP, Altaner CM, Apperley DC, Wess TJ & Jarvis MC (2013) Structure of cellulose microfibrils in primary cell walls from collenchyma. *Plant Physiol.* **161:** 465–476

Tone Y, Pedersen LC, Yamamoto T, Izumikawa T, Kitagawa H, Nishihara J, Tamura JI, Negishi M & Sugahara K (2008) 2-O-phosphorylation of xylose and 6-O-sulfation of galactose in the protein linkage region of glycosaminoglycans influence the glucuronyltransferase-I activity involved in the linkage region synthesis. *J. Biol. Chem.* **283:** 16801–16807

Tryfona T, Sorieul M, Feijao C, Stott K, Rubtsov D V., Anders N & Dupree P (2019) Development of an oligosaccharide library to characterise the structural variation in glucuronoarabinoxylan in the cell walls of vegetative tissues in grasses. *Biotechnol. Biofuels* **12:** 109

Tryfona T & Stephens E (2010) Analysis of carbohydrates on proteins by offline normal-phase liquid chromatography MALDI-TOF/TOF-MS/MS. In *Methods in Molecular Biology*, Cutillas PR & Timms JF (eds) pp 137–151. New York: Humana Press

Tu L & Banfield DK (2010) Localization of Golgi-resident glycosyltransferases. *Cell. Mol. Life Sci.* **67:** 29–41

Tu L, Tai WCS, Chen L & Banfield DK (2008) Signal-Mediated Dynamic Retention of Glycosyltransferases in the Golgi. *Science* **321:** 404–407

Ulvskov P, Paiva DS, Domozych D & Harholt J (2013) Classification, Naming and Evolutionary History of Glycosyltransferases from Sequenced Green and Red Algal Genomes. *PLoS One* **8:** e76511

Unterreitmeier S, Fuchs A, Schäffler T, Heym RG, Frishman D & Langosch D (2007) Phenylalanine Promotes Interaction of Transmembrane Domains via GxxxG Motifs. *J. Mol. Biol.* **374:** 705–718

Urbanowicz BR, Peña MJ, Moniz HA, Moremen KW & York WS (2014) Two Arabidopsis

proteins synthesize acetylated xylan in vitro. *Plant J.* **80:** 197–206

Valley CC, Lewis AK & Sachs JN (2017) Piecing it together: Unraveling the elusive structure-function relationship in single-pass membrane receptors. *Biochim. Biophys. Acta - Biomembr.* **1859:** 1398–1416

Vanbeselaere J, Jin C, Eckmair B, Wilson IBH & Paschinger K (2020) Sulfated and sialylated N-glycans in the echinoderm Holothuria atra reflect its marine habitat and phylogeny. *J. Biol. Chem.* **295:** 3159–3172

Varki A (2011) Evolutionary forces shaping the Golgi glycosylation machinery: Why cell surface glycans are universal to living cells. *Cold Spring Harb. Perspect. Biol.* **3:** 1–14

Varki A (2017) Biological roles of glycans. *Glycobiology* **27:** 3–49

Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, Stanley P, Hart G, Darvill A, Kinoshita T, Prestegard JJ, Schnaar RL, Freeze HH, Marth JD, Bertozzi CR, Etzler ME, Frank M, Vliegenthart JFG, Lütteke T, Perez S, et al (2015) Symbol nomenclature for graphical representations of glycans. *Glycobiology* **25:** 1323–1324

Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW & Etzler ME (2009) Essentials of Glycobiology, 2nd edition Cold Spring Harbor Laboratory Press

Vasconcelos A & Pomin V (2017) The Sea as a Rich Source of Structurally Unique Glycosaminoglycans and Mimetics. *Microorganisms* **5:** 51

Vilanova E, Santos GRC, Aquino RS, Valle-Delgado JJ, Anselmetti D, Fernàndez-Busquets X & Mourão PAS (2016) Carbohydrate-carbohydrate interactions mediated by sulfate esters and calcium provide the cell adhesion required for the emergence of early metazoans. *J. Biol. Chem.* **291:** 9425–9437

Vincken J-P, De Keizer A, Beldman C, Gerard A & Voragen J (1995) Fractionation of Xyloglucan Fragments and Their lnteraction with Cellulose. *Plant Physiol* **108:** 1579–1585

Vincken J-P, York WS, Beldman C & Voragen AGJ (1997) Two General Branching Patterns of Xyloglucan, XXXG and XXGG. *Plant Physiol.* **114:** 9–13

Volpi S, Yamazaki Y, Brauer PM, van Rooijen E, Hayashida A, Slavotinek A, Sun Kuehn H, Di Rocco M, Rivolta C, Bortolomai I, Du L, Felgentreff K, Ott de Bruin L, Hayashida K, Freedman G, Marcovecchio GE, Capuder K, Rath P, Luche N, Hagedorn EJ, et al (2017) EXTL3 mutations cause skeletal dysplasia, immune deficiency, and developmental delay. *J. Exp. Med.* **214:** 623–637

Wang J, Salem DR & Sani RK (2019) Extremophilic exopolysaccharides: A review and new perspectives on engineering strategies and applications. *Carbohydr. Polym.* **205:** 8–26

Wang M, Xu Z, Guo S, Zhou G, ONeill M & Kong Y (2020a) Identification of two functional xyloglucan galactosyltransferase homologs BrMUR3 and BoMUR3 in brassicaceous vegetables. *PeerJ* **8:** e9095

Wang S, Li L, Li H, Sahu SK, Wang H, Xu Y, Xian W, Song B, Liang H, Cheng S, Chang Y, Song Y, Çebi Z, Wittek S, Reder T, Peterson M, Yang H, Wang J, Melkonian B, Van de Peer Y, et al (2020b) Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat. Plants* **6:** 95–106

Wang Y, Ju T, Ding X, Xia B, Wang W, Xia L, He M & Cummings RD (2010a) Cosmc is an essential chaperone for correct protein O-glycosylation. *Proc. Natl. Acad. Sci. U. S. A.* **107:** 9228–9233

Wang Z, Ly M, Zhang F, Zhong W, Suen A, Hickey AM, Dordick JS & Linhardt RJ (2010b) E. coli K5 fermentation and the preparation of heparosan, a bioengineered heparin precursor. *Biotechnol. Bioeng.* **107:** 964–973

Weber E, Engler C, Gruetzner R, Werner S & Marillonnet S (2011) A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6:** e16765

Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, Xia E, Lu Y, Tai Y, She G, Sun J, Cao H, Tong W, Gao Q, Li Y, Deng W, Jiang X, Wang W, Chen Q, Zhang S, et al (2018) Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. U. S. A.* **115:** E4151–E4158

Wei G, Bai X, Gabb MMG, Bame KJ, Koshy TI, Speari PG & Esko JD (2000) Location of the glucuronosyltransferase domain in the heparan sulfate copolymerase EXT1 by analysis of Chinese hamster ovary cell mutants. *J. Biol. Chem.* **275:** 27733–27740

Weigel D & Glazebrook J (2006) Setting Up Arabidopsis Crosses. *Cold Spring Harb. Protoc.* **2006:** pdb.prot4623

Welch LG & Munro S (2019) A tale of short tails, through thick and thin: investigating the sorting mechanisms of Golgi enzymes. *FEBS Lett.* **593:** 2452–2465

Wen J, Xiao J, Rahdar M, Choudhury BP, Cui J, Taylor GS, Esko JD & Dixon JE (2014) Xylose phosphorylation functions as a molecular switch to regulate proteoglycan biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **111:** 15723–15728

Weng JK & Chapple C (2010) The origin and evolution of lignin biosynthesis. *New Phytol.* **187:** 273–285

Weston DJ, Turetsky MR, Johnson MG, Granath G, Lindo Z, Belyea LR, Rice SK, Hanson DT, Engelhardt KAM, Schmutz J, Dorrepaal E, Euskirchen ES, Stenøien HK, Szövényi P, Jackson M, Piatkowski BT, Muchero W, Norby RJ, Kostka JE, Glass JB, et al (2018)

The Sphagnome Project: enabling ecological and evolutionary insights through a genus-level sequencing project. *New Phytol.* **217:** 16–25

White FH (1961) Regeneration of Native Secondary and Tertiary Structures by Air Oxidation of Reduced Ribonuclease WITH A NOTE. *J. Biol. Chem.* **236:** 1353–1360

White SH & Wimley WC (1999) Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28:** 319–365

Wierzbicki MP, Maloney V, Mizrachi E & Myburg AA (2019) Xylan in the Middle: Understanding Xylan Biosynthesis and Its Metabolic Dependencies Toward Improving Wood Fiber for Industrial Processing. *Front. Plant Sci.* **10:** 176

Wiggins CAR & Munro S (1998) Activity of the yeast MNN1 α-1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proc. Natl. Acad. Sci. U. S. A.* **95:** 7945–7950

Wilson H V. (1907) On some phenomena of coalescence and regeneration in sponges. *J. Exp. Zool.* **5:** 245–258

Wrabl JO & Grishin N V. (2001) Homology between O-linked GlcNAc transferases and proteins of the glycogen phosphorylase superfamily. *J. Mol. Biol.* **314:** 365–374

Wu A-M, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P & Marchant A (2010a) Analysis of the Arabidopsis IRX9/IRX9-L and IRX14/IRX14-L pairs of glycosyltransferase genes reveals critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiol.* **153:** 542–54

Wu A-M, Rihouey C, Seveno M, Hörnblad E, Singh SK, Matsunaga T, Ishii T, Lerouge P & Marchant A (2009) The Arabidopsis IRX10 and IRX10-LIKE glycosyltransferases are critical for glucuronoxylan biosynthesis during secondary cell wall formation. *Plant J.* **57:** 718–731

Wu A, Hao P, Wei H, Sun H, Cheng S, Chen P, Ma Q, Gu L, Zhang M, Wang H & Yu S (2019) Genome-Wide Identification and Characterization of Glycosyltransferase Family 47 in Cotton. *Front. Genet.* **10:** 824

Wu AM, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P & Marchant A (2010b) Analysis of the arabidopsis IRX9/IRX9-L and IRX14/IRX14-L pairs of glycosyltransferase genes reveals critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiol.* **153:** 542–554

Wuyts W, Van Hul W, De Boulle K, Hendrickx J, Bakker E, Vanhoenacker F, Mollica F, Lüdecke HJ, Sayli BS, Pazzaglia UE, Mortier G, Hamel B, Conrad EU, Matsushita M, Raskind WH & Willems PJ (1998) Mutations in the EXT1 and EXT2 genes in hereditary

multiple exostoses. *Am. J. Hum. Genet.* **62:** 346–354

Xiao Y, Zeng B, Berner N, Frishman D, Langosch D & George Teese M (2020) Experimental determination and data-driven prediction of homotypic transmembrane domain interfaces. *Comput. Struct. Biotechnol. J.* **18:** 3230–3242

Xu H, Ding A, Chen S, Marowa P, Wang D, Chen M, Hu R, Kong Y, O'Neill M, Chai G & Zhou G (2018) Genome-wide analysis of sorghum GT47 family reveals functional divergences of MUR3-like genes. *Front. Plant Sci.* **9:** 1773

Yamada S (2020) Specific functions of Exostosin-like 3 (EXTL3) gene products. *Cell. Mol. Biol. Lett.* **25:** 39

Yamada S, Sugahara K & Özbek S (2011) Evolution of glycosaminoglycans: Comparative biochemical study. *Commun. Integr. Biol.* **4:** 150–158

Yamamoto F & Hakomori S (1990) Sugar-nucleotide donor specificity of histo-blood group A and B transferases is based on amino acid substitutions. *J. Biol. Chem.* **265:** 19257–19262

Yamamoto FI, Clausen H, White T, Marken J & Hakomori SI (1990) Molecular genetic basis of the histo-blood group ABO system. *Nature* **345:** 229–233

Yamamoto S & Oka S (2001) Glucuronyltransferases Involved in the HNK-1 Carbohydrate Epitope Biosynthesis. *Trends Glycosci. Glycotechnol.* **13:** 193–208

Yang C & Tang D (2000) Patient-Specific Carotid Plaque Progression Simulation. *C. Model. Eng. Sci.* **1:** 119–131

Yang J, Yan R, Roy A, Xu D, Poisson J & Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12:** 7–8

Yang M, Fehl C, Lees K V., Lim EK, Offen WA, Davies GJ, Bowles DJ, Davidson MG, Roberts SJ & Davis BG (2018) Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat. Chem. Biol.* **14:** 1109–1117

Yin Y, Mao X, Yang J, Chen X, Mao F & Xu Y (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40:** 445–451

York WS, van Halbeek H, Darvill AG & Albersheim P (1990) Structural analysis of xyloglucan oligosaccharides by 1H-n.m.r. spectroscopy and fast-atom-bombardment mass spectrometry. *Carbohydr. Res.* **200:** 9–31

York WS & O'Neill MA (2008) Biochemical control of xylan biosynthesis - which end is up? *Curr. Opin. Plant Biol.* **11:** 258–265

Yoshida-Moriguchi T & Campbell KP (2014) Matriglycan: A novel polysaccharide that links dystroglycan to the basement membrane. *Glycobiology* **25:** 702–713

Young WW (2004) Organization of Golgi Glycosyltransferases in Membranes: Complexity

via Complexes. *J. Membr. Biol.* **198:** 1–13

Yu L, Lyczakowski JJ, Pereira CS, Kotake T, Yu X, Li A, Mogelsvang S, Skaf MS & Dupree P (2018) The patterned structure of galactoglucomannan suggests it may bind to cellulose in seed mucilage. *Plant Physiol.* **178:** 1011–1026

Yu L, Lyczakowski JJ, Wightman R, Wilson LFL, Yoshimi Y, Yu X, Terrett OM, Stott K, Charalambous S, Wurman-Rodrich J, Temple H, Krogh KBRM & Dupree P (2021a) Structural and functional similarities between mannan and xyloglucan in the growing plant cell wall. Manuscript in preparation.

Yu L, Terrett OM, Wilson LFL, Wurman-Rodrich J, Lyczakowski JJ, Krogh KBRM & Dupree P (2021b) Identification of enzymes responsible for the addition of arabinose and galactose modifications onto glucuronic acid branches of xylan. Manuscript in preparation.

Yuan Z, Fang Y, Zhang T, Fei Z, Han F, Liu C, Liu M, Xiao W, Zhang W, Wu S, Zhang M, Ju Y, Xu H, Dai H, Liu Y, Chen Y, Wang L, Zhou J, Guan D, Yan M, et al (2018) The pomegranate (Punica granatum L.) genome provides insights into fruit quality and ovule developmental biology. *Plant Biotechnol. J.* **16:** 1363–1374

Zablackis E, York WS, Pauly M, Hantus S, Reiter WD, Chapple CCS, Albersheim P & Darvill A (1996) Substitution of L-fucose by L-galactose in cell walls of Arabidopsis mur1. *Science* **272:** 1808–1810

Zabotina OA (2012) Xyloglucan and Its Biosynthesis. *Front. Plant Sci.* **3:** 134

Zabotina OA, van de Ven WTG, Freshour G, Drakakaki G, Cavalier D, Mouille G, Hahn MG, Keegstra K & Raikhel N V. (2008) Arabidopsis *XXT5* gene encodes a putative α-1,6-xylosyltransferase that is involved in xyloglucan biosynthesis. *Plant J.* **56:** 101–115

Zak BM, Crawford BE & Esko JD (2002) Hereditary multiple exostoses and heparan sulfate polymerization. *Biochim. Biophys. Acta - Gen. Subj.* **1573:** 346–355

Zavyalov A V., Rykov S V., Lunina NA, Sushkova VI, Yarotsky S V. & Berezina O V. (2019) Plant Polysaccharide Xyloglucan and Enzymes That Hydrolyze It (Review). *Russ. J. Bioorganic Chem.* **45:** 845–859

Zeng W, Jiang N, Nadella R, Killen TL, Nadella V & Faik A (2010) A glucurono(arabino)xylan synthase complex from wheat contains members of the GT43, GT47, and GT75 families and functions cooperatively. *Plant Physiol.* **154:** 78–97

Zeng W, Lampugnani ER, Picard KL, Song L, Wu A-MM, Farion IM, Zhao J, Ford K, Doblin MS & Bacic A (2016) Asparagus IRX9, IRX10, and IRX14A Are Components of an Active Xylan Backbone Synthase Complex that Forms in the Golgi Apparatus. *Plant*

Physiol. **171:** 93–109

Zhang F, Zhang Z & Linhardt RJ (2010) Glycosaminoglycans. *Handb. Glycomics***:** 59–80

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y & Yin Y (2018a) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46:** W95–W101

Zhang H, Zhu F, Yang T, Ding L, Zhou M, Li J, Haslam SM, Dell A, Erlandsen H & Wu H (2014) The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. *Nat. Commun.* **5:** 4339

Zhang H, Zhu Q, Cui J, Wang Y, Chen MJ, Guo X, Tagliabracci VS, Dixon JE & Xiao J (2018b) Structure and evolution of the Fam20 kinases. *Nat. Commun.* **9:** 1218

Zhang J, Fu XX, Li RQ, Zhao X, Liu Y, Li MH, Zwaenepoel A, Ma H, Goffinet B, Guan YL, Xue JY, Liao YY, Wang QF, Wang QH, Wang JY, Zhang GQ, Wang ZW, Jia Y, Wang MZ, Dong SS, et al (2020a) The hornwort genome and early land plant evolution. *Nat. Plants* **6:** 107–118

Zhang K (2016) Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193:** 1–12

Zhang L, Chen F, Zhang X, Li Z, Zhao Y, Lohaus R, Chang X, Dong W, Ho SYW, Liu X, Song A, Chen J, Guo W, Wang Z, Zhuang Y, Wang H, Chen X, Hu J, Liu Y, Qin Y, et al (2020b) The water lily genome and the early evolution of flowering plants. *Nature* **577:** 79–84

Zhang Y, Skolnick J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I & Bourne P (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinforma. 2008 91* **59:** 305–309

Zheng SQ, Palovcak E, Armache JP, Verba KA, Cheng Y & Agard DA (2017) MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14:** 331–332

Zhong R, Cui D & Ye Z-H (2018) Secondary cell wall biosynthesis. *New Phytol.* **221:** 1703–1723

Zhong R, Peña MJ, Zhou GK, Nairn CJ, Wood-Jones A, Richardson EA, Morrison WH, Darvill AG, York WS & Ye ZH (2005) Arabidopsis fragile fiber8, which encodes a putative glucuronyltransferase, is essential for normal secondary wall synthesis. *Plant Cell* **17:** 3390–3408

Zhou G-K, Zhong R, Himmelsbach DS, McPhail BT & Ye Z-H (2007) Molecular Characterization of PoGT8D and PoGT43B, Two Secondary Wall-Associated Glycosyltransferases in Poplar. *Plant Cell Physiol.* **48:** 689–699

References

Zhu L, Dama M & Pauly M (2018) Identification of an arabinopyranosyltransferase from *Physcomitrella patens* involved in the synthesis of the hemicellulose xyloglucan. *Plant Direct* **2:** e00046

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN & Alva V (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430:** 2237–2243

Zivanov J, Nakane T, Forsberg BO, Kimanius D, Hagen WJH, Lindahl E & Scheres SHW (2018) New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7:** e42166

Zmora N, Suez J & Elinav E (2019) You are what you eat: diet, health and the gut microbiota. *Nat. Rev. Gastroenterol. Hepatol.* **16:** 35–56

Appendix : Bioinformatics scripts

Awk script for extracting entries from a FASTA file using a list of IDs (allowing for additional characters at the end of the ID):

```
awk 'FNR==NR{a[$0];next} /^>/{val=$0;sub(/^>/,"",val);flag=0;for \\
(val0 in a) if(match(val,val0)) {flag=1;break}} flag' ids.txt fasta_file
```

Python script `TMHgrab.py`, for running TMHMM and extracting the predicted TMH from the original sequence with 20 amino acids on either side (assumes TMHMM installed with 'tmhmm' in PATH variable):

```
from subprocess import Popen, PIPE
import sys

fastapath = str(sys.argv[1])
fastafile = open(fastapath, 'r')
fasta = fastafile.readlines()

process = Popen(['tmhmm', fastapath, '--short'], stdout = PIPE)
out = process.stdout.readlines()

entries = ['']*len(out)
names = [None]*len(out)

c = 0
for line in fasta:
    if '>' in line:
        names[c] = line.rstrip()
        c += 1
    else:
        entries[c-1] += line.rstrip()

c = 0
for i in out:
```

```
    topo = i.split('/t')[-1].split('=')[-1].split('-')
    if len(topo) > 1:
        start = int(topo[0][1:]) - 21
        finish = int(topo[1].split('o')[0].split('i')[0]) + 20
        if start < 0:
            start = 0
        if finish < 0:
            finish = 0
        print(names[c])
        print(entries[c][start:finish])
    c += 1

fastafile.close()
```

Python script `DomainExtract.py`, for truncating an alignment to a specified numerical range (requires each sequence to contain no line-breaks):

```
import sys
path = str(sys.argv[1])
start = int(sys.argv[2])
end = int(sys.argv[3]) - 1

f = open(path, 'r')
flines = f.readlines()

for line in flines:
    if line[0] == '>':
        print(line.rstrip())
    else:
        print(line[start:end])

f.close()
```