Evolution and Function of the Plant Cell Wall Synthesis-Related Glycosyltransferase Family 8[1][W][OA]

Yanbin Yin, Huiling Chen, Michael G. Hahn, Debra Mohnen*, and Ying Xu

Computational Systems Biology Laboratory and Institute of Bioinformatics (Y.Y., H.C., Y.X.), BioEnergy Science Center (Y.Y., M.G.H., D.M., Y.X.), Department of Plant Biology (M.G.H.), Complex Carbohydrate Research Center (M.G.H., D.M.), and Department of Biochemistry and Molecular Biology (D.M., Y.X.), University of Georgia, Athens, Georgia 30602

Carbohydrate-active enzyme glycosyltransferase family 8 (GT8) includes the plant galacturonosyltransferase1-related gene family of proven and putative α-galacturonosyltransferase (*GAUT*) and GAUT-like (*GATL*) genes. We computationally identified and investigated this family in 15 fully sequenced plant and green algal genomes and in the National Center for Biotechnology Information nonredundant protein database to determine the phylogenetic relatedness of the GAUTs and GATLs to other GT8 family members. The GT8 proteins fall into three well-delineated major classes. In addition to GAUTs and GATLs, known or predicted to be involved in plant cell wall biosynthesis, class I also includes a lower plant-specific GAUT and GATL-related (GATR) subfamily, two metazoan subfamilies, and proteins from other eukaryotes and cyanobacteria. Class II includes galactinol synthases and plant glycogenin-like starch initiation proteins that are not known to be directly involved in cell wall synthesis, as well as proteins from fungi, metazoans, viruses, and bacteria. Class III consists almost entirely of bacterial proteins that are lipooligo/polysaccharide α-galactosyltransferases and α-glucosyltransferases. Sequence motifs conserved across all GT8 subfamilies and those specific to plant cell wall-related GT8 subfamilies were identified and mapped onto a predicted GAUT1 protein structure. The tertiary structure prediction identified sequence motifs likely to represent key amino acids involved in catalysis, substrate binding, protein-protein interactions, and structural elements required for GAUT1 function. The results show that the GAUTs, GATLs, and GATRs have a different evolutionary origin than other plant GT8 genes, were likely acquired from an ancient cyanobacterium (*Synechococcus*) progenitor, and separate into unique subclades that may indicate functional specialization.

Plant cell walls are composed of three principal types of polysaccharides: cellulose, hemicellulose, and pectin. Studying the biosynthesis and degradation of these biopolymers is important because cell walls have multiple roles in plants, including providing structural support to cells and defense against pathogens, serving as cell-specific developmental and differentiation markers, and mediating or facilitating cell-cell communication. In addition to their important roles within plants, cell walls also have many economic uses in human and animal nutrition and as sources of natural textile fibers, paper and wood products, and components of fine chemicals and medicinal products. The study of the biosynthesis and biodegradation of plant cell walls has become even more significant because cell walls are the major components of biomass (Mohnen et al., 2008), which is the most promising renewable source for the production of biofuels and biomaterials (Ragauskas et al., 2006; Pauly and Keegstra, 2008). Analyses of fully sequenced plant genomes have revealed that they encode hundreds or even thousands of carbohydrate-active enzymes (CAZy; Henrissat et al., 2001; Yokoyama and Nishitani, 2004; Geisler-Lee et al., 2006). Most of these CAZy enzymes (Cantarel et al., 2009) are glycosyltransferases (GTs) or glycoside hydrolases, which are key players in plant cell wall biosynthesis and modification (Cosgrove, 2005).

The CAZy database is classified into 290 protein families (www.cazy.org; release of September 2008), of which 92 are GT families (Cantarel et al., 2009). A number of the GT families have been previously characterized to be involved in plant cell wall biosynthesis. For example, the GT2 family is known to include cellulose synthases and some hemicellulose backbone synthases (Lerouxel et al., 2006), such as mannan synthases (Dhugga et al., 2004; Liepman et al., 2005), putative xyloglucan synthases (Cocuron et al.,

[1] This work was supported by the U.S. Department of Energy (grant no. DE–PS02–06ER64304), the National Science Foundation (grant nos. NSF/DBI–0354771, NSF/DEB–0830024, NSF/ITR–IIS–0407204, NSF/DBI–0542119, NSF/CCF0621700, and NSF/MCB–0646109), and the U.S. Department of Agriculture (grant no. NRI–CREES–2006–35318–17301 and grant no. 2010–65115–20396 from the National Institute of Food and Agriculture). The BioEnergy Science Center was supported by the Office of Biological and Environmental Research in the Department of Energy Office of Science.

* Corresponding author; e-mail dmohnen@ccrc.uga.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Ying Xu (xyn@bmb.uga.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.110.154229

2007), and mixed linkage glucan synthases (Burton et al., 2006). With respect to the synthesis of xylan, a type of hemicellulose, four Arabidopsis (*Arabidopsis thaliana*) proteins from the GT43 family, irregular xylem 9 (IRX9), IRX14, IRX9-L, and IRX14-L, and two proteins from the GT47 family, IRX10 and IRX10-L, are candidates (York and O'Neill, 2008) for glucuronoxylan backbone synthases (Brown et al., 2007, 2009; Lee et al., 2007a; Peña et al., 2007; Wu et al., 2009). In addition, three proteins have been implicated in the synthesis of an oligosaccharide thought to act either as a primer or terminator in xylan synthesis (Peña et al., 2007): two from the GT8 family (IRX8/GAUT12 [Persson et al., 2007] and PARVUS/GATL1 [Brown et al., 2007; Lee et al., 2007b]) and one from the GT47 family (FRA8/IRX7 [Zhong et al., 2005]).

The GT families involved in the biosynthesis of pectins have been relatively less studied until recently. In 2006, a gene in CAZy family GT8 was shown to encode a functional homogalacturonan α-galacturonosyltransferase, GAUT1 (Sterling et al., 2006). *GAUT1* belongs to a 25-member gene family in Arabidopsis, the GAUT1-related gene family, that includes two distinct but closely related families, the galacturonosyltransferase (*GAUT*) genes and the galacturonosyltransferase-like (*GATL*) genes (Sterling et al., 2006). Another *GAUT* gene, *GAUT8/QUA1*, has been suggested to be involved in pectin and/or xylan synthesis, based on the phenotypes of plant lines carrying mutations in this gene (Bouton et al., 2002; Orfila et al., 2005). It has further been suggested that multiple members of the GT8 family are galacturonosyltransferases involved in pectin and/or xylan biosynthesis (Mohnen, 2008; Caffall and Mohnen, 2009; Caffall et al., 2009).

Aside from the 25 *GAUT* and *GATL* genes, Arabidopsis has 16 other family GT8 genes, according to the CAZy database, which do not seem to have the conserved sequence motifs found in GAUTs and GATLs: HxxGxxKPW and GLG (Sterling et al., 2006). Eight of these 16 genes are annotated as galactinol synthase (GolS) by The Arabidopsis Information Resource (TAIR; www.arabidopsis.org), and three of these AtGolS enzymes have been implicated in the synthesis of raffinose family oligosaccharides that are associated with stress tolerance (Taji et al., 2002). The other eight Arabidopsis GT8 genes are annotated as plant glycogenin-like starch initiation proteins (PGSIPs) in TAIR. PGSIPs have been proposed to be involved in the synthesis of primers necessary for starch biosynthesis (Chatterjee et al., 2005). Hence, the GT8 family is a protein family consisting of enzymes with very distinct proven and proposed functions. Indeed, a suggestion has been made to split the GT8 family into two groups (Sterling et al., 2006), namely, the cell wall biosynthesis-related genes (GAUTs and GATLs) and the non-cell wall synthesis-related genes (GolSs and PGSIPs).

We are interested in further defining the functions of the GAUT and GATL proteins in plants, in particular their role(s) in plant cell wall synthesis. The apparent disparate functions of the GT8 family (i.e. the GAUTs and GATLs as proven and putative plant cell wall polysaccharide biosynthetic α-galacturonosyltransferases, the eukaryotic GolSs as α-galactosyltransferases that synthesize the first step in the synthesis of the oligosaccharides stachyose and raffinose, the putative PGSIPs, and the large bacterial GT8 family of diverse α-glucosyltransferases and α-galactosyltransferases involved in lipopolysaccharide and lipooligosaccharide synthesis) indicate that the GT8 family members are involved in several unique types of glycoconjugate and glycan biosynthetic processes (Yin et al., 2010). This observation led us to ask whether any of the GT8 family members are sufficiently closely related to *GAUT* and *GATL* genes to be informative regarding GAUT or GATL biosynthetic function(s) and/or mechanism(s).

To investigate the relatedness of the members of the GT8 gene family, we carried out a detailed phylogenetic analysis of the entire GT8 family in 15 completely sequenced plant and green algal genomes (Table I) and also included GT8 proteins from the National Center for Biotechnology Information (NCBI) nonredundant protein database (NCBI-nr), the largest protein sequence database worldwide. Our specific goal was to define GT8 orthologs across different plant genomes and to investigate the evolutionary relationships among the different GT8 subfamilies. During this study, we also identified both conserved amino acid sequence motifs present in all GT8 subfamilies and those motifs specific to the GT8 subfamilies relevant to cell wall synthesis. Lastly, we analyzed these motifs in terms of their structural locations and possible functions based on a predicted GAUT1 protein tertiary (3D) structure to provide information that can be used to direct future GAUT and GATL structure-function studies.

RESULTS

In Silico Identification of GT8 Proteins

GT8 proteins in 15 genomes (Table I) and in the NCBI-nr database were identified using an HMMER search (Finn et al., 2006) for the Pfam Glyco_transf_8 (PF01501; 345 amino acids long) domain (referred to as the GT8 domain throughout this article; for details, see "Materials and Methods") as the query. This HMMER search, using an E-value cutoff of $\leq$1e-2, identified 378 nonidentical GT8 domains in 378 proteins from the 15 genomes, which are herein referred to as the 15genome-GT8 set. During the search, if two sequences of varying length but 100% identical over the shorter length were identified, only the longer one was kept. In addition, another 1,708 GT8 domains in 1,701 proteins were identified in the NCBI-nr database and termed the nr-GT8 set. The two data sets were combined and the redundant sequences were removed (for

Table I. *The nine plant and six green algal genomes used in this study*
JGI, Joint Genome Institute; TIGR, The Institute for Genomic Research.

Abbreviation	Clade	Species	Genome Published	Downloaded from
mpc	Green algae	*Micromonas pusilla* CCMP1545	Worden et al. (2009)	JGI version 2.0
mpr	Green algae	*Micromonas* strain RCC299	Worden et al. (2009)	JGI version 2.0
ol	Green algae	*Ostreococcus lucimarinus*	Palenik et al. (2007)	JGI version 1.0
ot	Green algae	*Ostreococcus tauri*	Derelle et al. (2006)	JGI version 1.0
cr	Green algae	*Chlamydomonas reinhardtii*	Merchant et al. (2007)	JGI version 3.0
vc	Green algae	*Volvox carteri* f. *nagariensis*	No	JGI version 1.0
pp	Moss	*Physcomitrella patens* ssp. *patens*	Rensing et al. (2008)	JGI version 1.1
sm	Spike moss	*Selaginella moellendorffii*	No	JGI version 1.0
pt	Dicot	*Populus trichocarpa*	Tuskan et al. (2006)	JGI version 1.1
at	Dicot	*Arabidopsis thaliana*	Arabidopsis Genome Initiative (2000)	TAIR version 9.0
vv	Dicot	*Vitis vinifera*	Jaillon et al. (2007)	http://www.genoscope.cns.fr/
gm	Dicot	*Glycine max*	Schmutz et al. (2010)	JGI version 1.0
os	Monocot	*Oryza sativa*	Goff et al. (2002); Yu et al. (2002)	TIGR version 6.1
sb	Monocot	*Sorghum bicolor*	Paterson et al. (2009)	JGI version 1.0
bd	Monocot	*Brachypodium distachyon*	Vogel et al. (2010)	JGI version 1.0

details, see "Materials and Methods") to generate a single set containing 918 nonredundant GT8 domain sequences, called the all-GT8 set. This set contains all of the GT8 genes in Arabidopsis identified previously (Sterling et al., 2006) and no non-GT8 genes. The 15genome-GT8 set and all-GT8 set and additional gene identifier information are available in Supplemental Table S1.

Phylogenetic Classification of Plant GT8 Proteins into Different Subfamilies

A phylogenetic tree for the 15genome-GT8 set was constructed based on the multiple sequence alignment of the full-length sequence of 378 GT8 proteins from plants. As shown in Figure 1, seven major monophyletic clusters form, representing seven different GT8 clades, six of which contain Arabidopsis genes that have been studied previously. Specifically, the GAUT clade contains 15 Arabidopsis GAUT proteins (Sterling et al., 2006), the GATL clade contains 10 Arabidopsis GATL proteins (Sterling et al., 2006), the GolS clade contains eight Arabidopsis GolS proteins (Taji et al., 2002), and the three PGSIP clades contain eight Arabidopsis PGSIP proteins (Chatterjee et al., 2005). Furthermore, several of the GT8 clades are split into multiple subclades. Specifically, the GAUT clade is subdivided into seven subclades, the GATL clade into at least five subclades, and the PGSIP-A clade into at least four subclades, based on the groupings within each clade. The number of proteins in each of the GT8 clades, calculated based on the groupings given in Figure 1, is shown in Table II.

The following novel observations can be made about this phylogenetic tree and the associated clade classifications. (1) The plant cell wall synthesis-related GAUT/GATL clades are phylogenetically distant from the other plant GT8 clades (i.e. GolS and PGSIP). (2) From lower plants to higher plants, there is a sharp increase in the number of GT8 genes, especially in the GAUT and GATL clades. (3) In total, 25 moss and 25 spike moss GT8 proteins (Table II) are found in the phylogeny (Fig. 1). Among them, four moss and two spike moss proteins form a monophyletic cluster, the seventh major GT8 clade of Figure 1, closest to the GAUT clade. However, the phylogenetic position of this lower plant-specific clade cannot be determined unambiguously from the phylogenetic analyses performed here, since its location varies depending on whether the full-length protein sequences (Fig. 1) or the GT8 domains (Supplemental Fig. S1) are used for the analyses (see below). We have given the name GAUT- and GATL-Related (GATR) to the proteins in this clade. The homology search against the NCBI-nr database also found another protein (GenBank gi: 116790957) from the conifer *Picea sitchensis* that clusters into the GATR clade. In addition, within the GAUT clade, proteins from either moss or spike moss or both are found in four of the seven GAUT subclades: the ones containing AtGAUT1 to AtGAUT3, AtGAUT8 and AtGAUT9, AtGAUT10 and AtGAUT11, and AtGAUT12 to AtGAUT15. Five moss proteins and one spike moss protein are found in the GATL clade, with the former placed basal to the rest of the GATL subclades. There are no *Physcomitrella* proteins in the GolS clade, which, however, does include four *Selaginella* proteins. Each of the three PGSIP clades contains both moss and spike moss proteins. (4) Two green algal GT8 proteins were identified. One of them, from *Ostreococcus lucimarinus* (GenBank gi: 145356270), is placed ancestral to the land plant GAUT and GATR clades. The other, from *Volvox carteri* f. *nagariensis* (not found in GenBank; JGI release identifier estExt_fgenesh4_pg.C_460009), is clustered with the PGSIP-B clade. (5) The eight Arabidopsis PGSIP proteins are grouped into three clades. The first one, PGSIP-A, includes the Arabidopsis PGSIP1 to PGSIP5 proteins. The PGSIP-B and PGSIP-C clades cluster with the GolS clade with a supporting value of 76%. Note that the FastTree program (Price et al., 2009), which was used

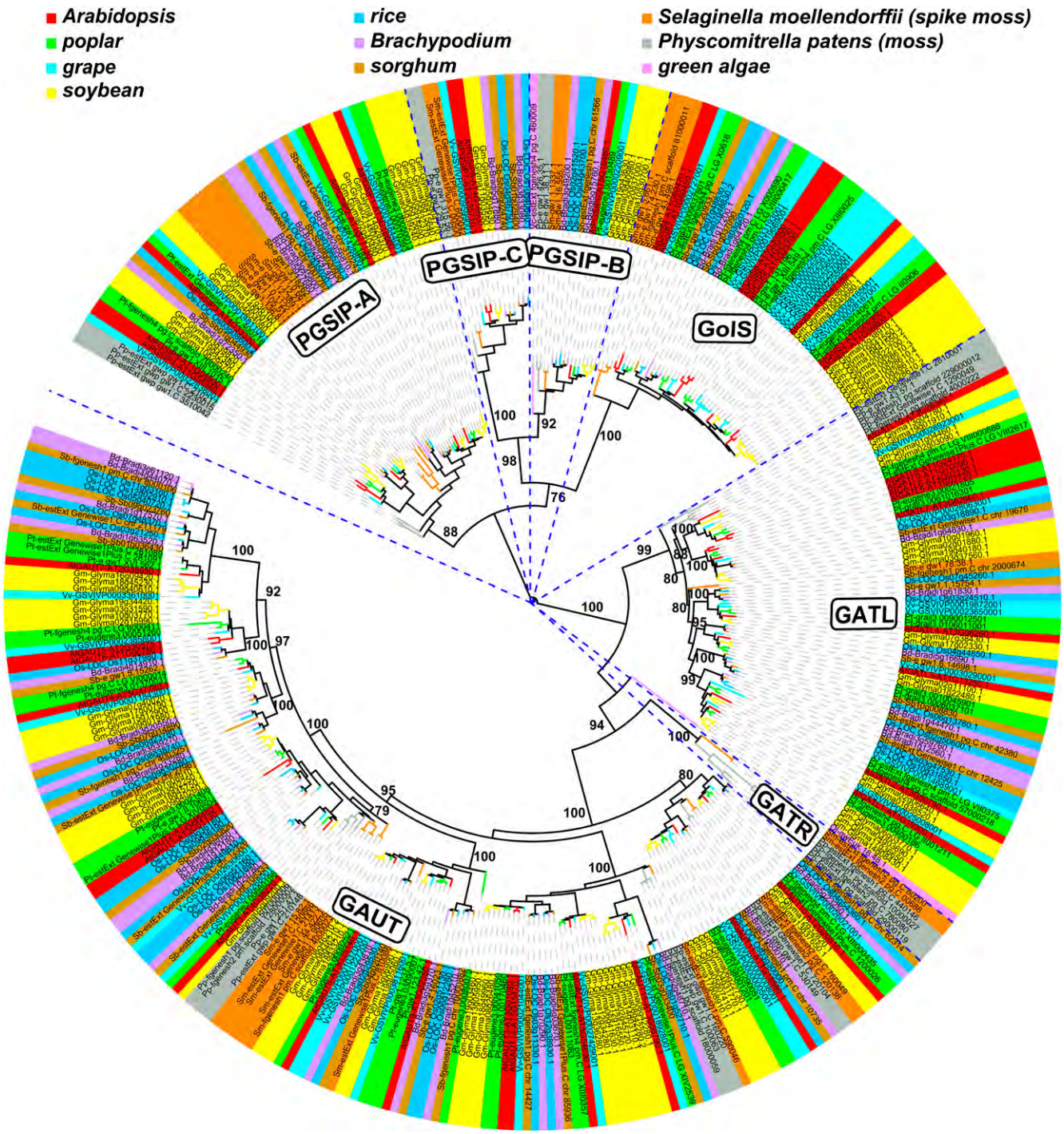


Figure 1. Phylogeny of 378 full-length GT8 proteins. The multiple sequence alignment of the full-length proteins was performed using the MAFFT version 6.603 program (Katoh et al., 2005) with the L-INS-I method. The phylogeny was reconstructed using the FastTree version 2.1.1 program (Price et al., 2009). The local support values beside the nodes were computed by resampling the site likelihoods 1,000 times and performing the Shimodaira Hasegawa test to show the confidence levels with regard to the clustering of relevant proteins into one group. Selected supporting values greater than 70% are shown. Major clades are identified by names: GAUT, GATL, GATR, GolS, and PGSIP. The phylogeny is displayed using the Interactive Tree of Life Web server (Letunic and Bork, 2007). An easier to view version of this figure is available online in the html version of the article.

to generate the phylogenies shown in Figures 1 and 2, tends to give very similar phylogenetic topology and classification but higher supporting values compared with traditional maximum likelihood programs like PhyML (Guindon and Gascuel, 2003), so that at least 80% is considered to be a "good" supporting value in

Table II. *The number of genes from different plant and algal genomes in different GT8 clades*

Abbreviation[a]	All	GAUT	GATL	GATR	GolS	PGSIP-A	PGSIP-B	PGSIP-C
mpc	0	0	0	0	0	0	0	0
mpr	0	0	0	0	0	0	0	0
ol	1	0	1	0	0	0	0	0
ot	0	0	0	0	0	0	0	0
cr	0	0	0	0	0	0	0	0
vc	1	0	0	0	0	0	1	0
pp	25	9	5	4	0	3	2	2
sm	25	10	1	2	4	5	1	2
pt	51	23	12	0	9	6	1	0
at	41	15	10	0	8	5	1	2
vv	37	14	7	0	10	4	1	1
gm	79[b]	37[c]	18	0	6[d]	12	4	2
os	39[e]	22	8	0	2[f]	3	2	2
sb	36	19	7	0	2	5	1	2
bd	36	19	6	0	2	5	2	2
Total	371[g]	168[h]	75	6	44[i]	48	16	15

[a]See Table I for full species names. [b-i]The numbers are 85, 39, 10, 40, 3, 378, 170, and 48, respectively, when alternatively spliced variants are included.

our experience. In this sense, the clustering of the GolS clade with the PGSIP-B and PGSIP-C clades is not significantly supported. The PGSIP-B clade includes the Arabidopsis PGSIP6 protein, and the PGSIP-C clade includes the Arabidopsis PGSIP7 and PGSIP8 proteins.

In addition to the full-length protein phylogeny, we also generated a phylogeny using the GT8 domains (Supplemental Fig. S1), which is the most highly conserved domain among all of the GT8 proteins examined. The full-length protein phylogeny (Fig. 1) has a topology very similar, but not identical, to the GT8 domain phylogeny (Supplemental Fig. S1). Noteworthy differences in the two topologies include the following. (1) The GATR clade clusters with the GAUT clade with a supporting value of 94% in the full-length protein phylogeny, while in the GT8 domain phylogeny, the GATR clade is basal to GAUT/GATL clades with 100% supporting value. (2) The GAUT/GATL-like green algal protein (GenBank gi: 145356270) clusters more with GAUT/GATR clades (supporting value of less than 70%) in the full-length phylogeny, while in the GT8 domain phylogeny, this green algal protein clusters with the GATL clade with a supporting value of 82%. (3) While the PGSIP-B and PGSIP-C clades cluster together with high statistical support in both phylogenies, the relationships of these two clades with the PGSIP-A and GolS clades remain somewhat ambiguous. In the full-length phylogeny, the PGSIP-B and PGSIP-C clades are clustered with the GolS clade, while in the GT8 domain phylogeny, the PGSIP-A clusters loosely with the GolS clade, with a modest grouping support value of 84%. (4) There are some subtle differences in subclade clustering between the full-length sequence and GT8 domain phylogenies. For example, in the full-length phylogeny, there is no monocot GAUT12 subcluster but rather two monocot subclusters in the GAUT13-14 subcluster. In contrast, in the GT8 domain phylogeny, there are single monocot subclusters in both the GAUT13-14 subcluster and in the GAUT12 subcluster.

Three Major GT8 Classes in Nature

In order to study the GT8 family in a broader evolutionary context, larger phylogenetic trees were generated for the all-GT8 set using both full-length protein sequences and Pfam GT8 domain regions from multiple organismal groups, namely plants, animals, fungi, bacteria, and viruses. The GT8 domain-based phylogeny is shown in Figure 2, while the full-length phylogeny is given in Supplemental Figure S2. We believe the domain phylogeny to be more accurate because the alignment quality is expected to be better for conserved domain regions than for full-length protein sequences (Ogden and Rosenberg, 2006; Talavera and Castresana, 2007), especially when aligning sequences from evolutionarily very distant organisms such as plants and bacteria. The unrooted tree constructed using the FastTree program (Price et al., 2009; Fig. 2) revealed that there are three distinct classes of GT8 proteins in nature.

The first class (I) contains the plant GAUT, GATL, and GATR clades and one of the green algal GT8 proteins (GenBank gi: 145356270), as shown in Figure 1. In addition, it also contains two metazoan GT8 clades, one of which, named Metazoan-1, is phylogenetically close to the plant GAUT/GATL/GATR clades, while the other, named Metazoan-2, is phylogenetically quite distant from the plant cell wall-related clades. Two other eukaryotic proteins in GT8 class I (shown in yellow-green) form a clade with the green algal protein (GenBank gi: 145356270). One of them (GenBank gi: 167521964) is from a marine choanoflagellate (*Monosiga brevicollis* MX1), the closest known unicellular relative of metazoans (King et al.,

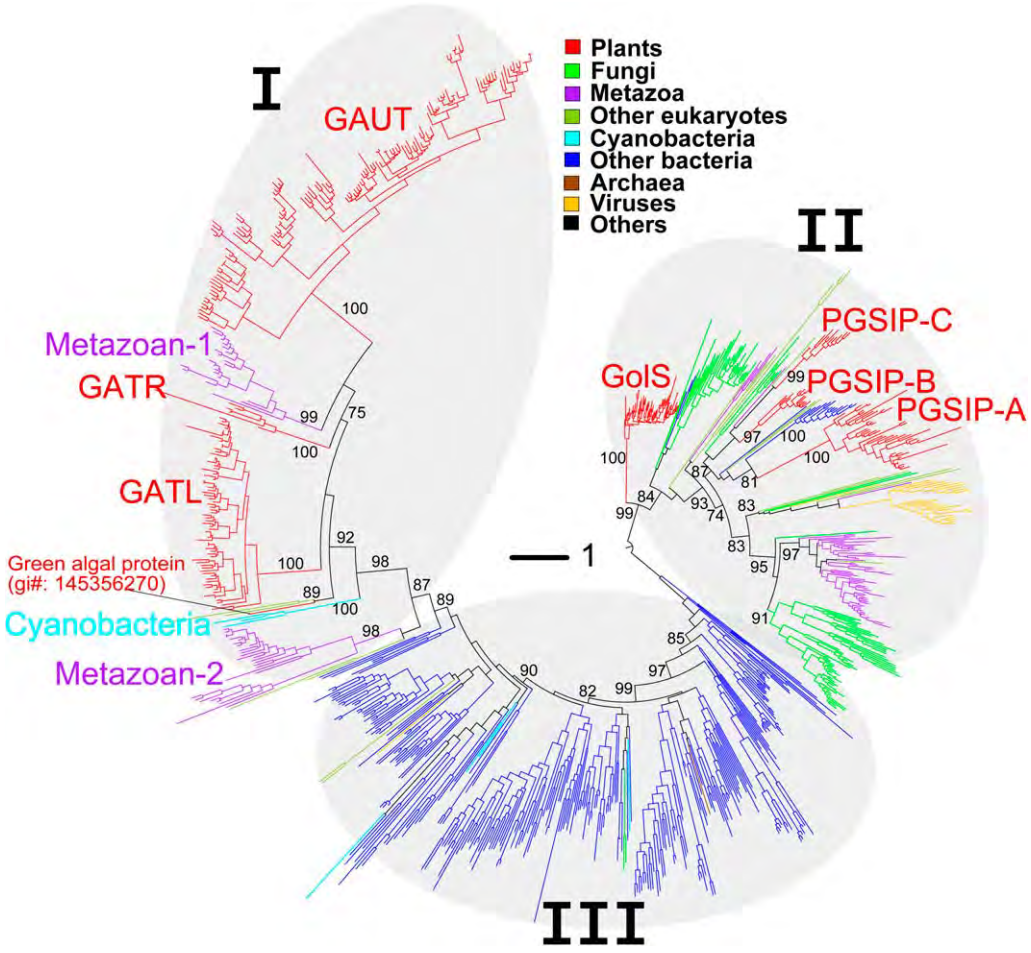


Figure 2. Phylogeny of the GT8 domains of 918 GT8 proteins. The multiple sequence alignment of the GT8 domains and the phylogeny reconstructions were performed as described in the legend to Figure 1. Major clades are identified and highlighted by their names to give clues about their function, species, or taxonomic information.

2008), while the other (GenBank gi: 239893977) is from a marine intracellular parasite (*Perkinsus marinus*). Close to the Metazoan-2 clade are two other eukaryotic proteins: one (GenBank gi: 219126835) is from *Phaeodactylum tricornutum CCAP 1055/1*, while the other (GenBank gi: 209877503) is from *Cryptosporidium muris* RN66. More interestingly, a clade of three cyanobacterial proteins (GenBank gi: 254421706, 254423034, and 81299339) from *Synechococcus* sp. PCC 7335 and *Synechococcus elongatus* PCC 7942 is placed at the base of the plant GAUT, GATL, GATR, and Metazoan-1 clades, with a very significant support value of 98%.

The second major class (II) contains the plant GolS and PGSIP clades as well as GT8 proteins from all the other organisms such as metazoans, fungi, viruses, and some bacteria. Note that the plant GolS and PGSIP proteins are unambiguously split into two distant clusters in this phylogeny, although they appeared to be more closely related in Figure 1. This is because the inclusion of more GT8 sequences from diverse organisms improved the phylogenetic resolution and, thus, effectively distinguished GolSs from PGSIPs.

The third class (III) of GT8 proteins consists almost exclusively of bacterial proteins. These proteins are from diverse types of bacteria, including Proteobacteria, Firmicutes, Cyanobacteria, Actinobacteria, and Bacteroidetes. A few fungal and viral proteins, possibly acquired horizontally from bacteria, are mixed together with the bacterial GT8 proteins, but no plant GT8 proteins are included in class III.

In summary, our phylogenetic analyses indicate that the GT8 family consists of three separate classes of proteins. Class I and class II contain largely eukaryotic proteins, while class III consists almost entirely of bacterial proteins. The plant cell wall-related proteins are all located in class I.

Rigorous Test of the Cyanobacterial Origin of Plant Cell Wall Biosynthesis-Related GT8 Genes

The phylogenetic analyses described above using the FastTree program (Price et al., 2009) placed a cyanobacterial clade basal to the plant cell wall-related GT8 proteins (class I; Fig. 2), suggesting that this

cyanobacterial clade might be a progenitor of the GAUT, GATL, and GATR proteins. To more rigorously test this hypothesis by phylogenetic analysis, a statistical test such as a bootstrap or Bayesian approach was needed to assess the statistical confidence in clustering the cyanobacterial proteins together with the plant cell wall-related GT8 proteins. However, it was not possible to use PhyML (Guindon and Gascuel, 2003) or MrBayes (Ronquist and Huelsenbeck, 2003), two of the most widely used accuracy-oriented phylogeny statistical tests, due to the excessive time and computer memory required to analyze so many sequences (918 proteins). Therefore, we selected representative sequences from the major clades observed in the tree (Fig. 2) and used these sequences to perform the phylogeny reconstruction and associated bootstrap and Bayesian analyses. Specifically, nine clades in classes I and II were selected: GAUT, GATL, GATR, Metazoan-1, Metazoan-2, GolS, and three PGSIP clades (for details, see "Materials and Methods"). These clades were selected to specifically test the cyanobacterial origin of plant class I GT8 proteins. Consensus sequences were built for each clade based on the multiple sequence alignment of GT8 domains in each of the nine selected clades. These consensus sequences were used as the representative sequences for their corresponding subfamilies. The reason that the consensus sequence was chosen to represent a clade is that a consensus sequence is a way of representing the multiple sequence alignment of all clade members: the amino acid at each position of the consensus sequence is the most abundant residue found in that column of the alignment, and the unconserved positions are shown as X. In another words, the consensus sequence captures the most conserved amino acid at each position of the clade alignment and therefore is most representative of the clade at each position.

An additional five single proteins from the GT8 tree (Fig. 2) were included in the data set for the rigorous statistical analyses: the green algal protein (GenBank gi: 145356270), the choanoflagellate protein (GenBank gi: 167521964), and one of the three cyanobacterial proteins (GenBank gi: 81299339) of class I and two GT8 proteins with solved 3D structures, LgtC from *Neisseria meningitidis* (Protein Data Bank [PDB] code 1GA8; Persson et al., 2001) and rabbit muscle glycogenin (PDB code 1LL2; Gibbons et al., 2002). The rabbit muscle glycogenin is from GT8 class II (Fig. 2) and is closely related to the plant PGSIP clades of class II. LgtC is a bacterial galactosyltransferase from GT8 class III. The selection of proteins included in this analysis covers all major clades in class I, since the goal was to assess the statistical support of the phylogenetic relationships of the proteins in class I. The analysis also included other plant GT8 clades in class II and some other proteins from the three classes.

The GT8 domains and the full-length proteins from the nine GT8 consensus sequences, together with those from the five individual GT8 proteins, were aligned. Based on this alignment, a maximum likelihood phylogeny with 100 replicate bootstrap analyses was built. In addition, a Bayesian approach was used to build the phylogeny and to assess the reliability of the resulting phylogenetic topology. The results of the phylogeny reconstructions, in which the class III protein LgtC from *N. meningitidis* was used as the outgroup, indicate that the topologies of the two phylogenies (i.e. the PhyML phylogeny and the MrBayes phylogeny) are largely similar, while the branch lengths and statistical supports for the internal nodes are different. The GT8 domain phylogeny (Fig. 3A) clearly shows that the cyanobacterial GT8 protein is ancestral to all class I GT8 clades, except for the Metazoan-2 clade, with good statistical support values by both PhyML and MrBayes analysis. Furthermore, when the phylogeny reconstruction was carried out using the full-length protein sequences instead of the Pfam GT8 domains, the statistical support for placing the cyanobacterial protein ancestral to all class I GT8 clades (except for the Metazoan-2 clade) remained significant (Fig. 3B). Thus, the statistical phylogenetic analyses strongly support the hypothesis that the *Synechococcus* cyanobacterial GT8 protein is directly ancestral to all other class I GT8 clades, except for Metazoan-2.

Identification of Conserved Sequence Motifs

Our phylogenetic analyses identified three major classes of GT8 proteins, each of which is divided into multiple clades containing multiple protein sequences. Proteins that carry out similar functions often contain conserved amino acid motifs. Therefore, we examined the protein alignments of the GT8 proteins to identify amino acid motifs that are conserved within and across clades, with particular emphasis on motifs that may relate the plant proteins in classes I and II to cell wall synthesis and provide insights into protein function. To this end, the multiple sequence alignments of the GT8 domains of the aforementioned 14 amino acid sequences, which contain the consensus sequences of nine GT8 clades and five individual GT8 sequences, were scanned for conserved amino acid domains. This was preferable to using the entire list of 918 GT8 sequences, because there is no straightforward way to display an alignment with hundreds of sequences that facilitates the identification of conserved motifs, and because each consensus sequence of the nine clades already captured all the sequence motifs of the clade in a simplified fashion. Therefore, the multiple sequence alignments of the consensus sequences were inspected to identify motifs that are conserved across all clades and to identify those motifs that are specific to plant cell wall-related clades.

Figure 4 shows the multiple sequence alignment, in which we also included the GT8 domain of the Arabidopsis GAUT1 protein (positions 334–647 of AtGAUT1/AT3G61130.1) as a reference, to locate the conserved motifs. Three secondary structures were also mapped onto the alignment: two were extracted

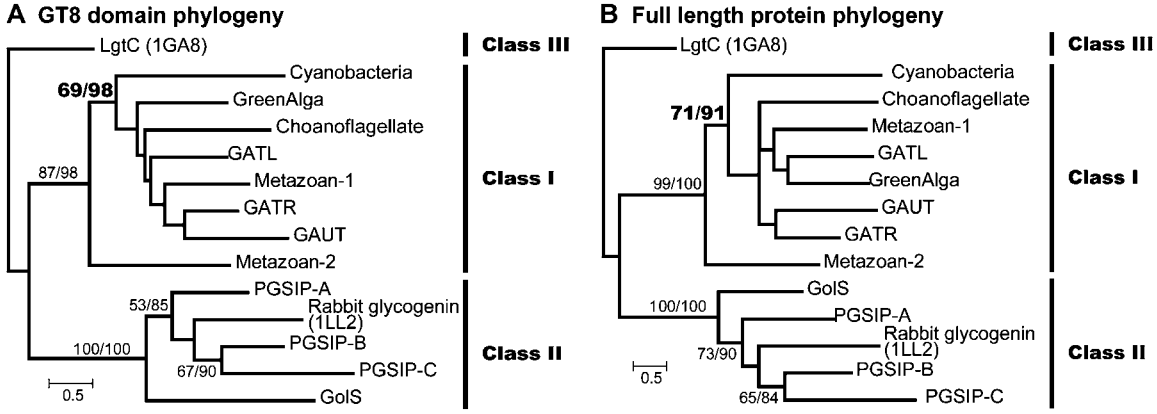


Figure 3. The maximum likelihood phylogeny of 14 representative GT8 amino acid sequences. LgtC, cyanobacteria, green alga, choanoflagellate, and rabbit glycogenin are five individual proteins taken from Figure 2, and the others are consensus sequences extracted from nine clades. A, The sequences used are the GT8 domain regions. B, The sequences used are the full-length proteins. The multiple sequence alignment of these sequences was used to build this phylogeny with 100 bootstrap analyses by PhyML version 3.0 (Guindon and Gascuel, 2003). Bayesian analysis was also performed with 1,000,000 generations using MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003). Bootstrap/Bayesian probability values of greater than 50% are shown beside each node.

from known GT8 PDB structures, LgtC (PDB code 1GA8; Persson et al., 2001) and rabbit muscle glycogenin (PDB code LL2; Gibbons et al., 2002), and one was extracted from a predicted 3D structure of AtGAUT1 (see below for details). Inspection of this alignment revealed both broadly conserved sequence domains and clade-specific sequence domains, highlighted in the alignment using triangles and stars, respectively. These conserved amino acid sequences include the DxD, HxxGxxKPW, and GLG motifs that were identified previously (Sterling et al., 2006) and three newly identified conserved amino acid sequence motifs (Fig. 5). Considering that each consensus sequence captures only the most frequent residue in each position, the original multiple sequence alignments of all of the individual proteins in the nine clades were inspected at each of the conserved sites to identify all of the amino acids in each column of the alignment in the conserved domains. This analysis, as summarized in Figures 4 and 5, identifies the DxD and HxxGxxKPW motifs as generally conserved across all GT8 clades, while the other four conserved motifs show both class-specific and clade-specific variations.

Prediction of the 3D Structure of AtGAUT1

The mapping of specifically identified sequence motifs onto the 3D structure of a relevant protein can, in some cases, provide insights into the possible functions of particular structural domains and amino acids within the protein. Toward this end, a structural prediction was done for AtGAUT1 using the LgtC structure 1GA8 (Persson et al., 2001) and the rabbit glycogenin structure 1LL2 (Gibbons et al., 2002) as the structural templates (for details, see "Materials and Methods"). Nine different protein-threading methods (Wu and Zhang, 2007) were used to identify the structurally similar templates of the AtGAUT1 GT8 domain (positions 334–647), and all nine programs indicated that the best template is 1GA8, a retaining α-1,4-galactosyltransferase (LgtC) involved in lipooligosaccharide synthesis in *N. meningitidis*, followed by 1LL2, a rabbit glycogenin, which are the only two solved GT8 structures in the PDB database. The predicted 3D structure of the C-terminal domain of AtGAUT1 (positions 367–647) superimposed on the 1GA8 structure of LgtC (Persson et al., 2001) is shown in Figure 6. There is a good match between the structures, with the exception of AtGAUT1 α-helix 3, LgtC α-helix 12, and the more flexible region near the acceptor binding site. The root mean square deviation between the two structures is 1.45 Å for 234 aligned residues (281 amino acids of GAUT1 and 271 amino acids of LgtC).

Mapping Sequence Motifs onto the Predicted AtGAUT1 3D Structure

The conserved amino acid motifs (Fig. 5) identified above were mapped onto the predicted structure of the AtGAUT1 GT8 domain, with the conserved amino acid positions highlighted using different colors by the ConSurf server (Landau et al., 2005). Examination of the predicted AtGAUT1 structure (Fig. 7A) using the information provided by the amino acid alignment (Fig. 4) reveals that most of the well-aligned regions are involved in forming a structural pocket in the protein structure (ligands in 1GA8 structure of LgtC are superimposed in Fig. 7, A and B). The root mean square deviation between the conserved binding pocket in the predicted structure for AtGAUT1 (Fig. 7B) and the structure of LgtC is 0.7 Å. These regions

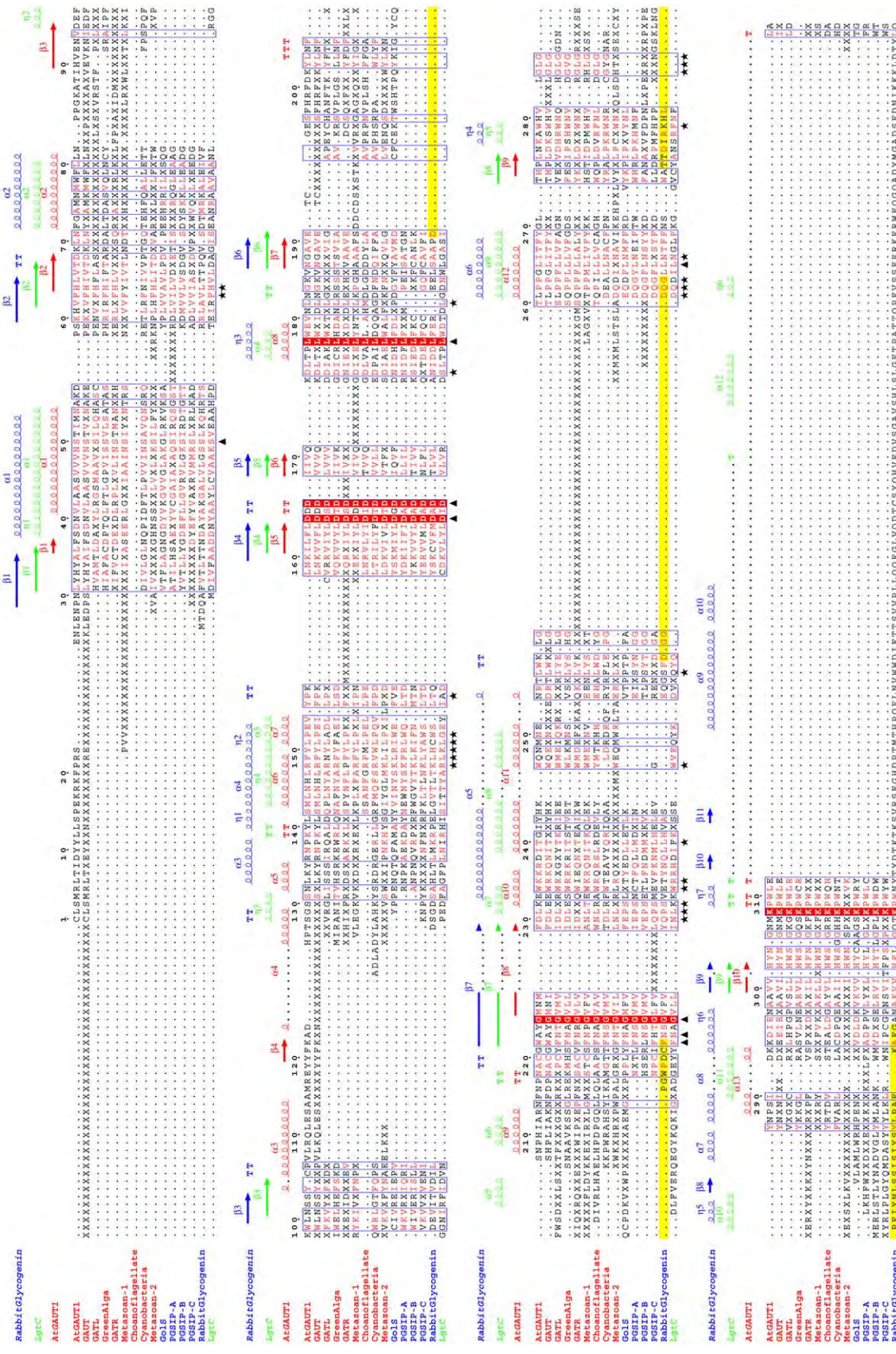
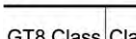


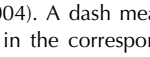
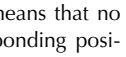
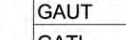
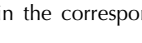
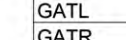
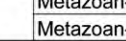
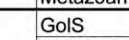
Figure 4. (*Legend appears on following page.*)

Figure 5. Conserved sequence motifs identified in the nine GT8 subfamilies of GT8 classes I and II. The positions of the motifs in the predicted secondary structure of AtGAUT1 are listed (see Fig. 4). A, The DxD and HxxGxxKPW motifs are relatively conserved across all clades. B, The motifs are distinct for class I (in some cases except for Metazoan-2) and class II. In both A and B, the first row shows the identified motifs, and rows 3 to 11 show the sequence logo of the motifs (extended in both ends in some cases) derived from the multiple sequence alignment of all members of each clade. The sequence logo was made using the WebLogo 3 server (Crooks et al., 2004). A dash means that no motif was found in the corresponding position.

A

GT8 Class	Clade	DxD	HxxGxxKPW
	Position	between a5 and b6	after b10
I	GAUT		
	GATL		
	GATR		
	Metazoan-1		
	Metazoan-2		
II	GolS		
	PGSIP-A		
	PGSIP-B		
	PGSIP-C		

B

GT8 Class	Clade	Wx(3,6)[G,H]LG/-	[L,Q,T]PP/ [D,E]Q[D,G]	[D,N]x(3)W/ [E,Q,K]P[S,N]	[R,P]xYL[P,A]/ Kx(3)[W,F]x(4)Y
	Position	b9 and a13	a12	b8 and a10	a6 and a7
I	GAUT				
	GATL				
	GATR				
	Metazoan-1				
	Metazoan-2				
II	GolS				
	PGSIP-A				
	PGSIP-B				
	PGSIP-C				

include residues 42 to 69 (α1 and β2), 144 to 191 (α6–7, β5–6, α8, and β7), 221 to 235 (β8 and α10), 255 to 268 (α12), and 265 to 277 (β10), using the amino acid coordinates of the AtGAUT1 GT8 domain. Furthermore, for amino acids in the pocket-forming helices, those inside the pocket tend to be more conserved than those outside the pocket (Fig. 7B). On the other hand, the weakly aligned regions, highlighted in the predicted structure with green and yellow colors (Fig. 7C), lie away from the structural pocket described above. It is noteworthy that the weakly aligned yellow regions in the predicted AtGAUT1 structure correspond to three segments in glycogenin (PDB code 1LL2; shaded with the yellow background in Figs. 4 and 7, A and C) that have been reported to be involved in dimer protein-protein interactions (Gibbons et al., 2002).

The conserved structural pocket on AtGAUT1 includes two generally conserved motifs in GT8 proteins that are located very close to each other (Fig. 7D): DxD and HxxGxxKPW (positions 165–167 and 302–310, respectively, of AtGAUT1 in Fig. 4). The close proximity of these two motifs suggests that there is coordination between the two motifs during the enzymatic reaction. Indeed, the DxD motif has been found to be conserved across diverse GT families (Wiggins and Munro, 1998) and is thought to be involved in interacting with Mn^{2+} and binding to nucleoside diphosphate (NDP)-sugar donors (Persson et al., 2001; Gibbons et al., 2002). The HxxGxxKPW motif had been identified previously by studying the alignment of 25 Arabidopsis GAUT and GATL sequences and is suggested to be part of the catalytic site of the enzyme (Sterling et al., 2006). Interestingly, our analyses (Figs. 4 and 5) indicate that this motif is present across most GT8 clades of classes I and II except for (1) the Metazoan-2 clade, for which only the H and K residues are conserved; (2) the PGSIP-C clade, for which only the KPW are conserved; and (3) the GolS clade, for which two additional amino acids are inserted between H and G. It has been suggested previously that

Figure 4. The multiple sequence alignment of 15 GT8 sequences. The 14 sequences from Figure 3 and the GT8 domain of AtGAUT1 were used to build this alignment using the MAFFT program (Katoh et al., 2005). The LgtC, cyanobacterium, green alga, choanoflagellate, and rabbit glycogenin are five individual proteins taken from Figure 2, while the others are consensus sequences extracted from nine clades as described in the text. X in the nine consensus sequences indicates a position that is not conserved in that clade. The graphical display of this alignment was generated, along with the added secondary structures above the alignment, using the ESPript Web server (Gouet et al., 2003). On the left, the sequence names in red, blue, and green indicate that they are from class I, II, and III, respectively. The regions shaded with yellow background in glycogenin (1LL2) are involved in forming a dimer complex (Gibbons et al., 2002). The most highly conserved sequence positions are marked with triangles at the bottom of the corresponding columns, while those that are different between class I (excluding Metazoan-2) and class II are marked with stars. The secondary structures indicated were extracted from the known GT8 PDB structures of LgtC (PDB code 1GA8; Persson et al., 2001) and glycogenin (PDB code 1LL2; Gibbons et al., 2002) or from a predicted 3D structure of AtGAUT1. Secondary structure symbols are as follows: β, β-sheet; α, α-helix; TT, strict β turns; TTT, strict α turns; η, ^{3}10 helices. The number associated with the secondary structure symbol represents the position of that type of secondary structure in each protein starting from the N terminus. The numbers associated with amino acid position are for the AtGAUT1 structure, with number 1 in the predicted structure being amino acid number 334 in the full-length protein.

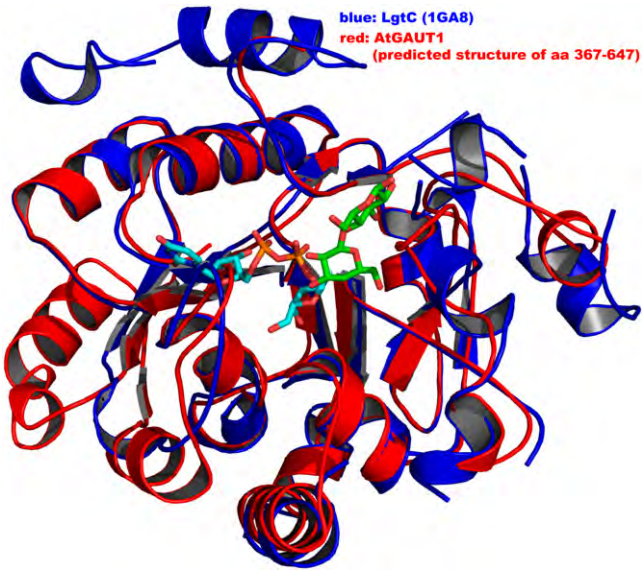


Figure 6. The predicted 3D structure of the GT8 domain of AtGAUT1 superimposed on the 1GA8 structure of LgtC. The predicted AtGAUT1 structure is shown in red, the structure of LgtC (Persson et al., 2001) is shown in blue, and the locations of the UDP-Gal substrate and the lactose portion of the acceptor for LgtC are shown in cyan and green, respectively. The GT8 domain of AtGAUT1 (amino acids 367–647) was used for the prediction using the I-TASSER prediction server (Zhang, 2008). PyMOL (http://www.pymol.org/) was used to generate the figure.

the H residue in HxxGxxKPW is involved in coordinating with Mn^{2+} and that the G and K are involved in interacting with the NDP-sugar donor (Persson et al., 2001; Gibbons et al., 2002). Both of these motifs are also found in the five individual proteins included in the detailed alignment analysis (Fig. 4), including the cyanobacterial protein (GenBank gi: 81299339) and the GATL-like green algal protein (GenBank gi: 145356270).

The GLG motif (positions 283–285 of AtGAUT1 in Fig. 4), which is different between class I and class II, was also identified previously (Sterling et al., 2006). Our analysis demonstrates that this motif is also present in the GATR clade. However, the GLG motif is modified to HLG in the Metazoan-1 clade, to GVG in the green algal protein, and to GYG in the cyanobacterial protein, while it is absent in the Metazoan-2 clade (Figs. 4 and 5). Unlike DxD and HxxGxxKPW, the GLG motif is exposed on the surface of the AtGAUT1 structure between β9 and α13, as highlighted using blue balls and sticks in Figure 7C.

The LPP motif is located at positions 261 to 263 of AtGAUT1 (Fig. 4). This motif is present across all GAUT and GATL proteins but is modified to QPP in the GATR clade and the green algal protein, to TPP in the Metazoan-1 clade, and to QPI in the choanoflagellate protein (Figs. 4 and 5). Interestingly, this motif is replaced by the DEA motif in the cyanobacterial protein and by the DQD motif in the Metazoan-2 clade of class I. Furthermore, there is an EQD motif in the corresponding positions in the GolS clade, DQG in the PGSIP clades and rabbit glycogenin, and DQD in LgtC of class III.

Two additional motifs are listed in Figure 5 that are different between class I and class II proteins and are also different between many of the clades in each class, suggesting that these motifs may be associated with clade-specific functions. In addition, there are even more amino acid positions highlighted with stars in Figure 4. All of these motifs and single positions are relatively less conserved and less discussed in the literature, but our multiple sequence alignment and structural mapping suggest that they are either involved in forming the binding pocket or presumably involved in other functions such as protein-protein interactions.

DISCUSSION

This study reports, to our knowledge, the first systematic and large-scale bioinformatic analysis of CAZy family GT8 proteins. Our study screened 15 completely sequenced plant and algal genomes, as well as the NCBI-nr database, to identify GT8 proteins. This comprehensive bioinformatic analysis of the GT8 family has provided novel observations about the GT8 family as a whole, identifying three distinct GT8 protein classes: class I, which includes the plant cell wall biosynthesis-related GAUT/GATL/GATR proteins; class II, which contains the GolS and PGSIP proteins; and class III, which is composed largely of bacterial GT8 proteins. The results presented here also identify a putative cyanobacterial progenitor to the GAUTs, GATLs, and GATRs. Finally, our analyses have identified several family-wide and clade-specific conserved sequence motifs whose locations in a predicted 3D structure of AtGAUT1 provide new insights into possible functional domains on this cell wall biosynthetic GT8 enzyme.

Plant GAUT/GATL/GATR Clades and GolS/PGSIP Clades Have Distinct Bacterial Origins

This large-scale phylogenetic analysis, which included both plant and nonplant GT8 proteins, revealed three major classes of GT8 proteins (Fig. 2), of which class III is composed almost exclusively of bacterial proteins. Class III also includes a small number of fungal and viral proteins and one Archaea protein (GenBank gi: 222444928), and it is possible that these organisms acquired their GT8 proteins from bacteria. The bacterial proteins in class III are from diverse bacterial phyla, including Proteobacteria, Firmicutes, Bacteroidetes, Actinobacteria, Cyanobacteria, and Spirochaetes, suggesting that GT8 is a very ancient glycosyltransferase family that evolved before all of these different bacterial phyla had diverged from one another, which may be bil-

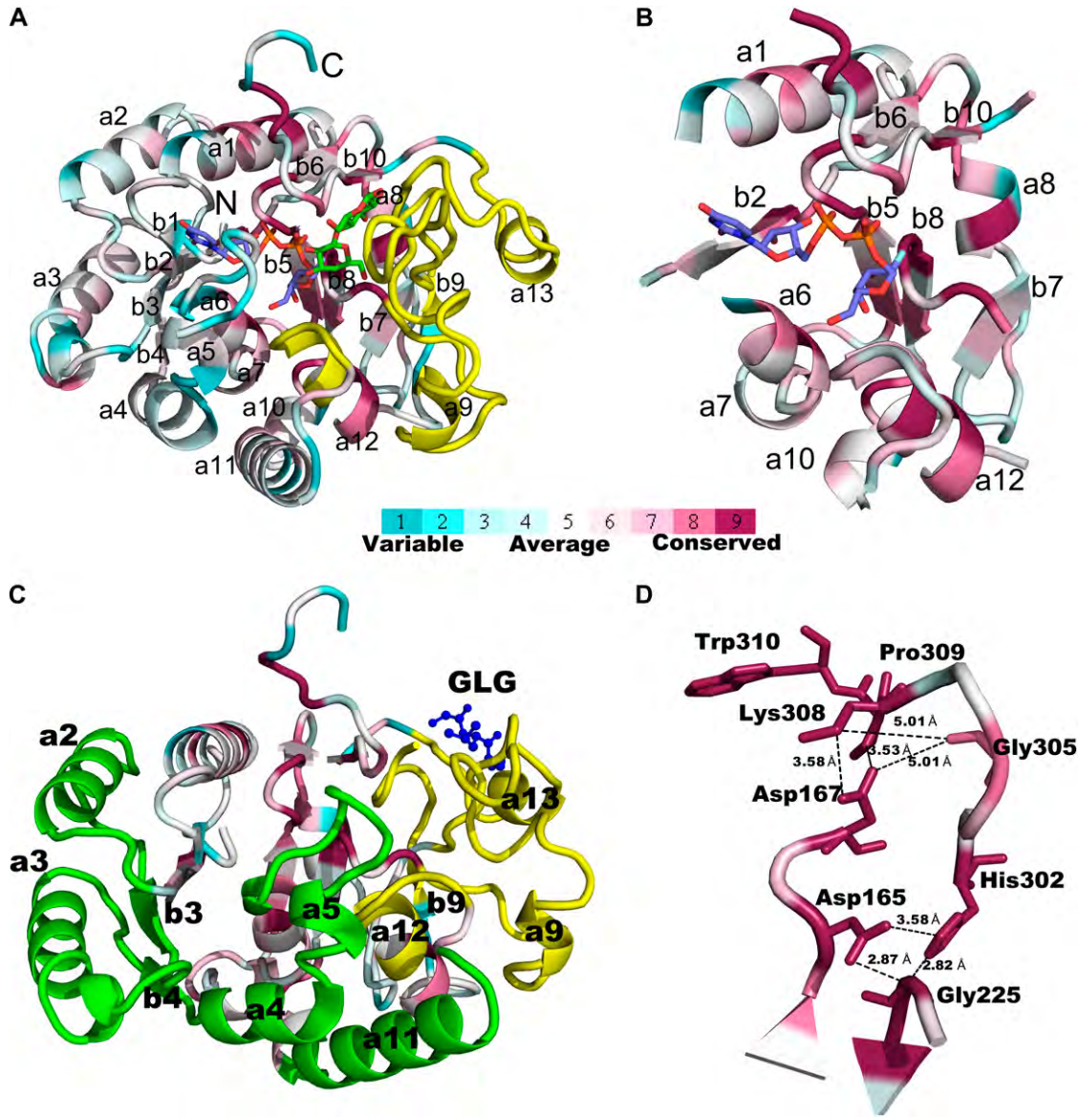


Figure 7. The predicted 3D structure of the GT8 domain of AtGAUT1. The GT8 domain of AtGAUT1 (amino acids 367–647) was used for the prediction as described in "Materials and Methods." The resulting PDB format file, the multiple sequence alignment shown in Figure 4, and a phylogeny inferred from the alignment were taken as inputs to the ConSurf server (Landau et al., 2005) to calculate a conservation score for each amino acid position in AtGAUT1. PyMOL (Delano, 2002) was used to generate the structures, with colors ranging from blue to red reflecting the conservation scores by the ConSurf prediction. A, The predicted 3D structure of AtGAUT1. The N and C termini are indicated. The UDP-Gal substrate and the lactose portion of the acceptor for LgtC are shown, in blue and green, respectively, in the binding pocket of the predicted structure. The yellow regions represent those that are shown with yellow background in glycogenin (1LL2) in Figure 4. The secondary structures of AtGAUT1 are highlighted with a for α-helix and b for β-sheet and are numbered from the N- to the C-terminal end of the threaded structure. B, The well-aligned regions shown in Figure 4 were extracted from the whole 3D structure. The binding pocket with the UDP-Gal substrate of LgtC is shown as described in A. The secondary structures of AtGAUT1 are indicated with a for α-helix and b for β-sheet and are numbered from the N- to the C-terminal end of the threaded structure. C, The weakly aligned regions are highlighted using green and yellow colors. The yellow regions represent those that are shown with yellow background in glycogenin (1LL2) in Figure 4. The GLG motif is highlighted in blue, and the relevant residues are displayed as balls and sticks. D, The conserved DxD motif and HxxGxxKPW motif are close to each other in the 3D space. The relevant key residues are displayed as sticks. The distances in the 3D space between residues are indicated.

lions of years ago (Cavalier-Smith, 2006). Hence, class III represents an ancestral GT8 gene pool from which the other two classes, consisting mostly of eukaryotic proteins, have evolved more recently. This is consistent with the ancient appearance of bacteria and the contributions of bacteria to the or-

igin of eukaryotes (Alberts et al., 2007; de Duve, 2007).

Interestingly, one cyanobacterial clade is clustered within class I and is basal to the plant cell wall synthesis-related GAUT/GATL/GATR clades and to the Metazoan-1 clade (Fig. 2). This placement of the cyanobacterial clade was confirmed by more rigorous phylogenetic analyses (Fig. 3) using two additional accuracy-oriented phylogeny reconstruction programs. This strongly suggests that an ancient cyanobacterium contributed to the origin of plant class I clades and is consistent with the proposal that cyanobacteria have played a major role in the origination of plants by an ancient endosymbiosis event that gave rise to chloroplasts (McFadden, 2001). The hypothesis of the cyanobacterial origin of plant cell wall-related GT8 genes is further supported by our sequence motif analysis, which shows that the cyanobacterial GT8 proteins share three class I-specific motifs (Fig. 5) that are distinct from motifs found in classes II and III (Fig. 4).

The basal position of the cyanobacterial clade within GT8 class I has fundamental significance with regard to the origin of plant cell walls. It has been shown that cyanobacterial cellulose synthases are evolutionarily closer to plant cellulose synthases and to some hemicellulose backbone synthases (e.g. the cellulose synthase-like/CSL families) than to the other bacterial cellulose synthases (Nobles et al., 2001; Nobles and Brown, 2004). Together with our findings reported here, this suggests that the key progenitor enzymes responsible for the biosynthesis of the backbones of all three of the major classes of plant cell wall polysaccharides (i.e. celluloses, hemicelluloses, and pectins) were in existence anciently in the bacterial world and that at least some of them were subsequently transferred into plants, possibly at, or after the time of, the establishment of plastids (or cyanobacterial endosymbionts) in the ancestral plant.

It is noteworthy that the other plant GT8 clades, that is, the GolS and PGSIP clades, reside in the distinct class II along with clades from other eukaryotes, viruses, and bacteria (Fig. 2). The bacterial clade in class II is composed of proteins exclusively from Alphaproteobacteria and clusters with the plant PGSIP-A clade. This suggests that PGSIP-A may have been acquired by plants from Alphaproteobacteria anciently via a horizontal gene-transfer event, which conforms to the well-documented hypothesis that some Alphaproteobacteria are the progenitor of mitochondria (Gray et al., 1999) and contributed genes to the ancient eukaryotes through endosymbiotic gene transfer (Timmis et al., 2004). However, the possibility cannot be ruled out that ancient plants have contributed genes to Alphaproteobacteria via horizontal gene transfer in the other direction. The observation that the three plant PGSIP clades cluster with different nonplant clades suggests that they diverged very anciently and may have distinct origins.

Our phylogenetic analyses also provide direct phylogenetic evidence to support splitting or subdividing the GT8 family into distinct groupings according to the three major classes identified in this study (Figs. 1 and 2; Supplemental Figs. S1 and S2), as has been suggested previously (Sterling et al., 2006). In particular, the two classes (I and II) that contain plant GT8 proteins have distinct evolutionary origins from the bacteria-dominated class III, as noted above.

The Clade and Subclade Diversification of the Plant GT8s in Classes I and II Occurred Prior to the Emergence of Angiosperms

The positioning of green algal and moss proteins on the GT8 phylogenetic tree suggests that a significant part of the diversification of plant GT8 proteins occurred before the appearance of angiosperms. Thus, the clustering of two green algal proteins with the GAUT/GATL and PGSIP-B clades, respectively, suggests that these algal proteins represent the ancestral orthologs of all land plant GAUT/GATL and PGSIP-B proteins, respectively. This, in turn, suggests that, although the GAUT and GATL clades are evolutionarily closer to each other than to the GolS/PGSIP clades, they very likely diverged before land plants appeared.

The phylogenetic analyses also suggest that subclades observed within most of plant clades in class I and class II diverged before the emergence of angiosperms. For example, as shown in Figure 1, the GAUT clade is composed of seven subclades, each of which is supported by a significant supporting value (greater than 80%), except for the clade containing AtGAUT1 to AtGAUT3 (79%). Four of these subclades (those containing AtGAUT1 to AtGAUT3, AtGAUT8, AtGAUT10 and AtGAUT11, and AtGAUT12 to AtGAUT15) include proteins from moss or spike moss or both. Furthermore, the moss GAUT proteins tend to be located at basal positions in each of the four GAUT subclades in which they occur. Similarly, there are five moss GATLs that are positioned basally to all of the higher plant GATL clades. The basal positioning of these moss proteins in these GAUT and GATL clades and subclades suggests that the diversification of these protein families took place early in the evolution of land plants. Lastly, there are moss proteins positioned basally with respect to each of the three PGSIP clades and the GolS clade within class II, again supporting an early divergence of these protein families in plant evolution. Furthermore, the presence of basal moss proteins in the phylogeny suggests that the proteins in the three PGSIP clades might have divergent functions. Interestingly, AtPGSIP1 and AtPGSIP3 have been found to be coexpressed with the cellulose synthase gene IRX3/CesA7, which is involved in secondary cell wall synthesis (Brown et al., 2005), raising the possibility that at least some members of the PGSIP-A clade may be directly or indirectly involved in plant cell wall biosynthesis.

Plant GT8 Proteins Are Involved in Diverse Biological Processes with the GAUTs, GATLs, GATRs, and One Green Algae Protein Grouped into the Distinct Plant Cell Wall-Related Class I

Our phylogenetic analyses (Fig. 1; Supplemental Fig. S1) identified seven plant GT8 clades: the GAUT, GATL, GATR, GolS, and three PGSIP clades. The different plant GT8 clades have been shown or implicated to be involved in various polysaccharide and glycoconjugate biosynthetic processes in plants. For example, some GAUTs are proven (Sterling et al., 2006) and putative pectin α-galacturonosyltransferases (Bouton et al., 2002; Orfila et al., 2005; Persson et al., 2007; Caffall et al., 2009). Other GAUTs and GATLs are proposed to be involved in plant cell wall-related polysaccharide (pectin and xylan) synthesis, although for most, the specific enzyme activity remains to be determined (Brown et al., 2007; Lee et al., 2007b). In contrast, the PGSIP genes are proposed to be α-glucosyltransferases involved in the initiation of starch synthesis, which entails both protein and oligosaccharide glycosylation (Chatterjee et al., 2005), while the GolS genes are α-galactosyltransferases that add Gal onto myoinositol to initiate raffinose and stachyose synthesis (Taji et al., 2002). Thus, the acceptors used by the plant GT8 proteins vary widely, as do the types of nucleotide-sugar donor substrates (i.e. GalA, Gal, or Glc). However, the α-anomeric configuration and the retaining enzyme mechanism appear to be conserved in the family.

An original goal of this study was to determine the relatedness of the non-GAUT/GATL GT8 family clades to the GAUT and GATL genes in order to identify any closely related GT8 family members that might have glycan or glycoconjugate biosynthetic function(s) and/or mechanism(s) applicable to the GAUTs/GATLs. Such information was deemed useful due to the large diversity in substrate and acceptor specificities and glycan biosynthetic strategies used by the members of diverse GT8 clades. The results of this study show that the GAUT and GATL proteins fall into a distinct GT8 family class (class I) that is clearly distinguishable from the GolS and PGSIP clades (class II) and from the bacterial GT8 proteins (class III). This finding reduces the likelihood that a high degree of similarity in acceptor-substrate and/or overall biosynthetic strategies exists between the proteins in these distinct GT8 classes.

The GATR clade, which contains moss proteins and at least one green algae protein, was found to be phylogenetically close to the GAUT/GATL clades (Fig. 1; Supplemental Fig. S1). The close evolutionary relatedness of the GATR and green algal proteins to the GAUT and GATL clades, which are involved in the synthesis of pectins and possibly other wall polysaccharides, is reasonable since pectic polysaccharides and GalA have been found in both mosses and green algae (both Charophytes and Chlorophytes), at least at low concentrations (Popper and Fry, 2003; Popper, 2008). The sharp increase in the numbers of GAUTs and GATLs from lower plants to higher plants suggests that these clades have important roles in plants as they adjusted to living on land and adopting an upright growth habit with increased size.

Functional Implications from Structural Analyses of Conserved Sequence Motifs Mapped onto the Predicted AtGAUT1 3D Structure

Our motif search (Figs. 4 and 5) and structural analysis of motifs in the context of the predicted AtGAUT1 structure (Fig. 7) provide several novel insights into the functions of plant cell wall synthesis-related GT8 proteins. For example, the GLG motif (Figs. 4, 5, and 7C) and its modified version, HLG, are found in all class I clades except for Metazoan-2, and this motif is predicted to reside on the surface of the AtGAUT1 protein (Fig. 7). In addition, a W residue, three or six amino acids upstream from the [G,H]LG motif, is found to be highly conserved across almost all class I GT8 proteins (Figs. 4 and 5). In contrast, this motif is completely missing in proteins of classes II and III (Figs. 4 and 5). Interestingly, the corresponding region in glycogenin (1LL2) of class II is involved in protein-protein interactions during dimer formation (Gibbons et al., 2002), suggesting that the GLG motif might play a similar functional role that is specific to all of the plant class I GT8 proteins, including the GAUTs/GATLs.

Another example is the helix (α12) that intrudes into the presumed substrate-binding pocket of AtGAUT1 (Figs. 6 and 7; Supplemental Fig. S3). This helix contains the [L,Q,T]PP motif that is conserved in many clades of class I, including the GAUT/GATL clades (Fig. 4). The corresponding motif is DQG and DQD in rabbit glycogenin (Gibbons et al., 2002) and LgtC (Persson et al., 2001), respectively. In both solved glycogenin and LgtC PDB structures, this motif is believed to be involved in the interaction of the protein with the sugar group of the NDP-sugar ligand. Specifically, the side chain oxygen of Q (Gln) has been proposed to be a nucleophilic acceptor involved in formation of the glycosyl-enzyme intermediate, thereby temporally accepting the monosaccharide from the NDP-sugar donor and subsequently transferring it to the final oligosaccharide acceptor (Persson et al., 2001; Gibbons et al., 2002). Although a role for Q as a nucleophile is unusual, the arguments and data in support of such a role in LgtC are strong (Persson et al., 2001). It would be unexpected that P or L in the similarly positioned LPP motif in AtGAUT1 would function as a nucleophile.

Interestingly, however, there is a Thr (T) immediately N terminal to the L in the LPP motif of AtGAUT1 (Fig. 4). Thr, like Gln, is uncharged but polar. This T is the closest residue to the sugar group of the donor ligand in 3D space and is conserved in all GAUTs except for GAUT5, GAUT7, and GAUT15. The corresponding position in the GATLs is occupied by S,

which is generally functionally similar to T and is completely conserved in the GATLs. It is tempting to speculate that this T or S may serve as the nucleophilic acceptor in the GAUTs or GATLs. However, if this is true, then it would be unlikely that GAUT5, GAUT7, or GAUT15 would be catalytically active, since they contain G, A, or A, respectively, in the comparable position. In this respect, it is noteworthy that attempts to recover galacturonosyltransferase activity from heterologously expressed GAUT7, expressed under conditions that allow GAUT1 activity, have been unsuccessful (Sterling et al., 2006), suggesting that GAUT7 itself may not be catalytically active. There have been no reports on the enzyme activity of GAUT5 or GAUT15; thus, whether or not these GAUTs are catalytically active remains to be determined. The conservation of the [L,Q,T]PP motif in many class I proteins and its location in the 3D structure suggest that it plays some important role(s) in maintaining the integrity of the binding pocket and/or in catalysis. Future experimental studies are clearly needed to determine the role, if any, of this motif in class I enzymatic function.

GAUT1 catalyzes the addition of GalA from UDP-GalA onto growing α-1,4-linked homogalacturonan oligosaccharide and polysaccharide acceptors (Sterling et al., 2006; Mohnen, 2008). Thus, it might be expected that the acceptor-binding pocket would be lined with positively charged amino acids. Inspection of the electrostatic potential surface of the predicted GAUT1 structure does indeed identify an extensive patch of positively charged amino acids in the vicinity of where the acceptor may be expected to lie (Supplemental Fig. S3). However, the exact location of the homogalacturonan acceptor, which is significantly different from the lactosyl moiety of the LgtC lipooligosaccharide acceptor that is shown superimposed on the GAUT1 structure in Figures 6 and 7 and Supplemental Figure S3, remains to be determined. Furthermore, this region of the GAUT1 structure is flexible and also is the same region known to be involved in dimer formation in glycogenin (1LL2; Gibbons et al., 2002). An understanding of the position and mechanism whereby GAUT1 binds acceptor oligogalacturonides or larger homogalacturonan stretches awaits further experimentation.

CONCLUSION

In summary, this study makes three major contributions to our understanding of CAZy family GT8. (1) We provide strong phylogenetic evidence to support a subdivision of the GT8 family into three classes, with the cell wall biosynthesis-related GAUTs/GATLs in class I, the GolSs and PGSIPs in class II, and with class III composed largely of bacterial GT8 proteins. (2) We identified a *Synechococcus* cyanobacterial GT8 protein as a putative progenitor for all GAUTs and GATLs. (3) We identified conserved sequence motifs and mapped them onto the predicted 3D structure of AtGAUT1, providing a wealth of information for use in forming hypotheses regarding specific amino acid functions in GAUT1 catalysis and substrate binding. Our computational study has improved our understanding of this important cell wall biosynthesis-related enzyme family in terms of both evolution and function, and provides a firm foundation for future hypotheses and experimental studies directed toward delineating the structure and function of diverse GAUTs and GATLs as well as other GT8 family members.

MATERIALS AND METHODS

Data Sources

The 15 target genomes, predicted genes, and translated protein data were downloaded from various sources, as specified in Table I. The NCBI-nr database was downloaded at ftp://ftp.ncbi.nih.gov/blast/db/FASTA/ as of December 12, 2009. The 41 Arabidopsis (*Arabidopsis thaliana*) GT8 gene accession numbers were obtained from http://cellwall.genomics.purdue.edu/families/2-3-1.html (Yong et al., 2005). Sequences from the 41 proteins were either collected from Sterling et al. (2006) or from the CAZy database (Coutinho et al., 2003).

Homolog Search

We ran the hmmsearch command of the HMMER package (Eddy, 1998) by querying the Pfam Glyco_transf_8 (PF01501) Hidden Markov Model (HMM) in ls mode (global with respect to query domain and local with respect to hit protein; for details, see the manual of the HMMER package) against the 15 fully sequenced plant and algal genomes (their protein data set) as well as against the NCBI-nr database. An E-value cutoff of $\leq$1e-2 was adopted to identify significant protein matches. Using this cutoff, all previously known Arabidopsis proteins were identified and no Arabidopsis false-positive proteins were found (Sterling et al., 2006; Penning et al., 2009).

The CD-HIT program (Li and Godzik, 2006) was used to remove redundant protein sequences in a data set. If any two proteins in the data set were found to be 90% identical, only the longer one was kept (for details, see the user guide of CD-HIT).

Phylogenetic Analysis

For all analyses, multiple protein sequence alignments were performed using the MAFFT version 6.603 program (Katoh et al., 2005), employing L-INS-I, which is considered to be one of the most accurate multiple protein sequence alignment methods available (Ahola et al., 2006; Nuin et al., 2006). Unless otherwise indicated, phylogenies shown in this article were constructed using the FastTree version 2.1.1 program (Price et al., 2009), which implements an ultrafast and fairly accurate approximate maximum likelihood method. The accuracy of FastTree version 2.1.1 phylogeny is considered to be slightly better than PhyML version 3.0, with minimum-evolution nearest-neighbor interchanges moves, and is 100 to 1,000 times faster and requires much less computer memory (http://www.microbesonline.org/fasttree/). FastTree analyses were conducted with default parameters; specifically, the amino acid substitution matrix was JTT, the number of rate categories of sites (CAT model) was 20, and the local support values of each node were computed by resampling the site likelihoods 1,000 times and performing the Shimodaira Hasegawa test.

Selected phylogenies were reconstructed using either maximum likelihood or Bayesian methods. Maximum likelihood phylogenies were built using PhyML version 3.0 (Guindon and Gascuel, 2003). Specifically, PhyML analyses were conducted with the JTT model, 100 replicates of bootstrap analyses, estimated proportion of invariable sites, four rate categories, estimated gamma distribution parameter, and optimized starting BIONJ tree. Bayesian phylogenies were built using MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003). MrBayes analyses were conducted with a mixed amino acid model estimated in the run, estimated proportion of invariable sites, estimated

gamma distribution parameter, and 1,000,000 generations. In addition to maximum likelihood and Bayesian phylogenetic analyses, there are other phylogeny reconstruction algorithms, including neighbor-joining, maximum parsimony, and unweighted pair group method with arithmetic mean. However, it is generally known that these algorithms are less accurate than maximum likelihood and Bayesian analyses (Hall, 2005; Ogden and Rosenberg, 2006); thus, they were not used in this study.

Construction of Consensus Sequences for Selected Protein Subfamilies

Consensus sequences were calculated for nine GT8 subfamilies (GAUT, GATL, GATR, Metazoan-1, Metazoan-2, GolS, and three PGSIP clades) by running the cons command in the EMBOSS version 2.8.0 package, with the multiple sequence alignment of all protein members of each clade as the input.

Structural Prediction

The I-TASSER server (Zhang, 2008) was used to predict the protein structure for AtGAUT1. I-TASSER is one of the best protein structure prediction servers according to the recent worldwide experiments on critical assessment of techniques for protein structure prediction (Kopp et al., 2007; Cozzetto et al., 2009). The submitted sequence (the Pfam domain region of AtGAUT1, amino acid residues 334–647 of AT3G61130.1) first underwent multiple threading at LOMETS (Wu and Zhang, 2007), a meta-threading sever with nine locally installed threading programs, to identify structurally similar templates from PDB. It was found that PDB proteins LgtC (1GA8; Persson et al., 2001) and glycogenin (1LL2; Gibbons et al., 2002) were always the top two structural templates for the queried AtGAUT1 protein. Consensus target-template alignments from these threading programs were then collected and submitted for iterative structure refinement. An initial threaded AtGAUT1 structure was generated using amino acids 334 to 367. However, it was determined that the N-terminal 33 amino acids of the AtGAUT1 pfam domain constituted a loop region that did not align to the LgtC (1GA8) template in the structural alignment. Thus, the 33-amino acid N-terminal loop region of the AtGAUT1 pfam region was removed and the threading was redone using amino acids 367 to 647, to yield the AtGAUT1 structures shown in Figures 6 and 7 and Supplemental Figure S3. The two threaded structures were compared, and it was confirmed that there were no differences between the AtGAUT1 334 to 367 and 367 to 647 threaded structures, except for the N-terminal loop. Final atomic details were built through optimization of the hydrogen-bonding network.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Phylogeny of the Pfam GT8 domains of 378 plant and algal GT8 proteins.

Supplemental Figure S2. Phylogeny of 918 full-length GT8 proteins.

Supplemental Figure S3. Electrostatic surface of the predicted structure of AtGAUT1.

Supplemental Table S1. Part a, 378 plant GT8 proteins from 15 plant genomes; part b, 918 nonredundant GT8 family protein sequences.

ACKNOWLEDGMENTS

We acknowledge the U.S. Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) for the sequence data of *Selaginella moellendorffii*, *Volvox carteri* f. *nagariensis*, *Aureococcus anophagefferens*, and *Phaeodactylum tricornutum*, which were released prior to publication. We thank Dr. Igor Jouline and Dr. Brian Cantwell (Oak Ridge National Laboratories) for discussions of phylogenetic analyses and Dr. Harry Gilbert (Complex Carbohydrate Research Center) for discussions about 3D structure alignments. We are also grateful to the members of our three laboratories for discussions and comments.

Received February 1, 2010; accepted June 1, 2010; published June 3, 2010.

LITERATURE CITED

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E (2006) A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics **7:** 484

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2007) Genetic information in eucaryotes. *In* Molecular Biology of the Cell. Garland Science, New York, pp 26–30

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Bouton S, Leboeuf E, Mouille G, Leydecker MT, Talbotec J, Granier F, Lahaye M, Hofte H, Truong HN (2002) QUASIMODO1 encodes a putative membrane-bound glycosyltransferase required for normal pectin synthesis and cell adhesion in *Arabidopsis*. Plant Cell **14:** 2577–2590

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 LA, 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

Burton RA, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Medhurst A, Stone BA, Newbigin EJ, Bacic A, Fincher GB (2006) Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science **311:** 1940–1942

Caffall KH, Mohnen D (2009) The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. Carbohydr Res **344:** 1879–1900

Caffall KH, Pattathil S, Phillips SE, Hahn MG, Mohnen D (2009) *Arabidopsis thaliana* T-DNA mutants implicate GAUT genes in the biosynthesis of pectin and xylan in cell walls and seed testa. Mol Plant **2:** 1000–1014

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res **37:** D233–D238

Cavalier-Smith T (2006) Cell evolution and Earth history: stasis and revolution. Philos Trans R Soc Lond B Biol Sci **361:** 969–1006

Chatterjee M, Berbezy P, Vyas D, Coates S, Barsby T (2005) Reduced expression of a protein homologous to glycogenin leads to reduction of starch content in *Arabidopsis* leaves. Plant Sci **168:** 501–509

Cocuron JC, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG (2007) A gene from the cellulose synthase-like C family encodes a β-1,4 glucan synthase. Proc Natl Acad Sci USA **104:** 8550–8555

Cosgrove DJ (2005) Growth of the plant cell wall. Nat Rev Mol Cell Biol **6:** 850–861

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

Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins (Suppl 9) **77:** 18–28

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res **14:** 1188–1190

de Duve C (2007) The origin of eukaryotes: a reappraisal. Nat Rev Genet **8:** 395–403

Delano WL (2002) The PyMOL Molecular Graphics System. http://www.pymol.org (January 5, 2008)

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, et al (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc Natl Acad Sci USA **103:** 11647–11652

Dhugga KS, Barreiro R, Whitten B, Stecca K, Hazebroek J, Randhawa GS, Dolan M, Kinney AJ, Tomes D, Nichols S, et al (2004) Guar seed beta-mannan synthase is a member of the cellulose synthase super gene family. Science **303:** 363–366

Eddy SR (1998) Profile hidden Markov models. Bioinformatics **14:** 755–763

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al (2006) Pfam: clans, Web tools and services. Nucleic Acids Res **34:** D247–D251

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, et al (2006) Poplar carbohydrate-active enzymes: gene identification and expression analyses. Plant Physiol **140:** 946–962

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

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100

Gouet P, Robert X, Courcelle E (2003) ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. Nucleic Acids Res **31:** 3320–3323

Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. Science **283:** 1476–1481

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:** 696–704

Hall BG (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol Biol Evol **22:** 792–802

Henrissat B, Coutinho PM, Davies GJ (2001) A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. Plant Mol Biol **47:** 55–72

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res **33:** 511–518

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature **451:** 783–788

Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins (Suppl 8) **69:** 38–56

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res **33:** W299–W302

Lee C, O'Neill MA, Tsumuraya Y, Darvill AG, Ye ZH (2007a) The *irregular xylem9* mutant is deficient in xylan xylosyltransferase activity. Plant Cell Physiol **48:** 1624–1634

Lee CH, Zhong RQ, Richardson EA, Himmelsbach DS, McPhail BT, Ye ZH (2007b) The *PARVUS* gene is expressed in cells undergoing secondary wall thickening and is essential for glucuronoxylan biosynthesis. Plant Cell Physiol **48:** 1659–1672

Lerouxel O, Cavalier DM, Liepman AH, Keegstra K (2006) Biosynthesis of plant cell wall polysaccharides—a complex process. Curr Opin Plant Biol **9:** 621–630

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:** 127–128

Li WZ, Godzik A (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:** 1658–1659

Liepman AH, Wilkerson CG, Keegstra K (2005) Expression of cellulose synthase-like (Csl) genes in insect cells reveals that CslA family members encode mannan synthases. Proc Natl Acad Sci USA **102:** 2221–2226

McFadden GI (2001) Primary and secondary endosymbiosis and the origin of plastids. J Phycol **37:** 951–959

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, et al (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245–250

Mohnen D (2008) Pectin structure and biosynthesis. Curr Opin Plant Biol **11:** 266–277

Mohnen D, Bar-Peled M, Somerville CR (2008) Cell wall polysaccharide synthesis. *In* ME Himmel, ed, Biomass Recalcitrance: Deconstructing the Plant Cell Wall for Bioenergy. Blackwell Publishing, Singapore, pp 94–159

Nobles DR, Brown RM (2004) The pivotal role of cyanobacteria in the evolution of cellulose synthases and cellulose synthase-like proteins. Cellulose **11:** 437–448

Nobles DR, Romanovicz DK, Brown RM Jr (2001) Cellulose in cyanobacteria: origin of vascular plant cellulose synthase? Plant Physiol **127:** 529–542

Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics **7:** 471

Ogden TH, Rosenberg MS (2006) Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol **55:** 314–328

Orfila C, Sørensen SO, Harholt J, Geshi N, Crombie H, Truong HN, Reid JS, Knox JP, Scheller HV (2005) *QUASIMODO1* is expressed in vascular tissue of *Arabidopsis thaliana* inflorescence stems, and affects homogalacturonan and xylan biosynthesis. Planta **222:** 613–622

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, et al (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci USA **104:** 7705–7710

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551–556

Pauly M, Keegstra K (2008) Cell-wall carbohydrates and their modification as a resource for biofuels. Plant J **54:** 559–568

Peña MJ, Zhong R, Zhou GK, Richardson EA, O'Neill MA, Darvill AG, York WS, Ye ZH (2007) *Arabidopsis irregular xylem8* and *irregular xylem9*: implications for the complexity of glucuronoxylan biosynthesis. Plant Cell **19:** 549–563

Penning BW, Hunter CT III, Tayengwa R, Eveland AL, Dugard CK, Olek AT, Vermerris W, Koch KE, McCarty DR, Davis MF, et al (2009) Genetic resources for maize cell wall biology. Plant Physiol **151:** 1703–1728

Persson K, Ly HD, Dieckelmann M, Wakarchuk WW, Withers SG, Strynadka NCJ (2001) Crystal structure of the retaining galactosyltransferase LgtC from *Neisseria meningitidis* in complex with donor and acceptor sugar analogs. Nat Struct Biol **8:** 166–175

Persson S, Caffall KH, Freshour G, Hilley MT, Bauer S, Poindexter P, Hahn MG, Mohnen D, Somerville C (2007) The *Arabidopsis irregular xylem8* mutant is deficient in glucuronoxylan and homogalacturonan, which are essential for secondary cell wall integrity. Plant Cell **19:** 237–255

Popper ZA (2008) Evolution and diversity of green plant cell walls. Curr Opin Plant Biol **11:** 286–292

Popper ZA, Fry SC (2003) Primary cell wall composition of bryophytes and charophytes. Ann Bot (Lond) **91:** 1–12

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol **26:** 1641–1650

Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ Jr, Hallett JP, Leak DJ, Liotta CL, et al (2006) The path forward for biofuels and biomaterials. Science **311:** 484–489

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, et al (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64–69

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572–1574

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463:** 178–183

Sterling JD, Atmodjo MA, Inwood SE, Kumar Kolli VS, Quigley HF, Hahn MG, Mohnen D (2006) Functional identification of an *Arabidopsis* pectin biosynthetic homogalacturonan galacturonosyltransferase. Proc Natl Acad Sci USA **103:** 5236–5241

Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. Plant J **29:** 417–426

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol **56:** 564–577

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet **5:** 123–135

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, et al (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463:** 763–768

Wiggins CA, 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 USA **95:** 7945–7950

Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, et al (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science **324:** 268–272

Wu AM, 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 S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res **35:** 3375–3382

Yin Y, Mohnen D, Gelineo-Albersheim I, Xu Y, Hahn MG (2010) Glycosyltransferases of family 8 (GT8). *In* P Ulvskov, ed, Plant Cell Wall Polysaccharides: Biosynthesis and Bioengineering. Annual Plant Reviews. Blackwell Publishing, Singapore (in press)

Yokoyama R, Nishitani K (2004) Genomic basis for cell-wall diversity in plants: a comparative approach to gene families in rice and Arabidopsis. Plant Cell Physiol **45:** 1111–1121

Yong W, Link B, O'Malley R, Tewari J, Hunter CT, Lu CA, Li X, Bleecker AB, Koch KE, McCann MC, et al (2005) Genomics of plant cell wall biogenesis. Planta **221:** 747–751

York WS, O'Neill MA (2008) Biochemical control of xylan biosynthesis: which end is up? Curr Opin Plant Biol **11:** 258–265

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics **9:** 40

Zhong RQ, 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